

**SIMULATION EXPERIMENT PLATFORM FOR EVALUATING  
CLINICAL TRIAL DESIGNS, WITH APPLICATIONS TO PHASE I DOSE-FINDING  
CLINICAL TRIALS**

by

**Yuanyuan Wang**

BS, University of Science and Technology of China, China, 2002

MA, Johns Hopkins University, 2005

Submitted to the Graduate Faculty of  
Graduate School of Public Health in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

**Yuanyuan Wang**

It was defended on

**September 9th, 2010**

and approved by

Dissertation Advisor:

Roger Day, ScD

Associate Professor

Departments of Biomedical Informatics and Biostatistics

School of Medicine and Graduate School of Public Health

University of Pittsburgh

Committee Member:

Daniel Normolle, PhD

Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Committee Member:

Hussein Tawbi, MD

Assistant Professor

Department of Medicine

School of Medicine

University of Pittsburgh

Committee Member:

Abdus Wahed, PhD

Associate Professor

Department of Biostatistics

Graduate School of Public Health

University of Pittsburgh

Copyright © by Yuanyuan Wang

2010

SIMULATION EXPERIMENT PLATFORM FOR EVALUATING  
CLINICAL TRIAL DESIGNS, WITH APPLICATIONS TO PHASE I DOSE-FINDING  
CLINICAL TRIALS

Yuanyuan Wang, PhD

University of Pittsburgh, 2010

Clinical Trial (CT) simulation is used by academic research centers and pharmaceutical companies to improve the efficiency and accuracy of drug development. Sophisticated commercial software for CT simulations is available for those with resources to cover fees and with design challenges that happen to match the software's capabilities. Academic research centers usually use locally developed or shared software for study design, mainly due to cost and flexibility considerations. Inspired by the success and immense influence of open-source software development projects, we are building an open-source simulation experiment platform with the intention of utilizing the power of distributed study design expertise, development talent, and peer review of code. The code base relies on S4 classes and methods within R. Design, baseline characteristic model, population model, outcome model, and evaluation criterion are five key object types. An action queue-based approach allows for complex decision making at the patient or CT level. Name matching mechanism is used to check interoperability among the objects. Extensibility, reuse and sharing come from the class/method architecture, together with automatic object and documentation discovery mechanisms.

An extensive literature review of existing design evaluation criteria did not reveal the use of criteria based on utility functions. In this dissertation, we propose flexible criteria for

evaluating Phase I trial designs by assessing through CT simulation the expected total personal utility, societal utility and total utility.

To illustrate the application, we present several examples using the platform to investigate important questions in clinical trial designs. Specifically, we look at the logit model in the continual reassessment method (CRM), choices of parameterization and prior distribution for its model parameters, and the effect of patient heterogeneity on the performance of the standard “3+3” design and the CRM.

This work creates an open-source highly flexible and extensible platform for evaluating CT designs via simulation, and promotes collaborative statistical software development. Its impact on public health will manifest itself in greatly speeding and expanding thorough and thoughtful design evaluations when developing clinical trials, for a community of CT designers.

## TABLE OF CONTENTS

<b>ACKNOWLEDGEMENTS .....</b>	<b>XIV</b>
<b>1.0 INTRODUCTION.....</b>	<b>1</b>
<b>1.1 STATEMENT OF PROBLEMS .....</b>	<b>1</b>
<b>1.2 OBJECTIVES OF THE STUDY .....</b>	<b>4</b>
<b>2.0 OVERVIEW OF PHASE I DOSE-FINDING CLINICAL TRIAL DESIGNS.....</b>	<b>5</b>
<b>2.1 INTRODUCTION .....</b>	<b>5</b>
<b>2.2 RECENT DEVELOPMENT .....</b>	<b>9</b>
<b>2.2.1 Designs for single outcome.....</b>	<b>9</b>
<b>2.2.2 Designs for bivariate outcome .....</b>	<b>14</b>
<b>2.2.3 Summary .....</b>	<b>16</b>
<b>3.0 EVALUATION CRITERIA FOR PHASE I DOSE-FINDING CLINICAL TRIAL DESIGNS.....</b>	<b>17</b>
<b>3.1 EXISTING EVALUATION CRITERIA.....</b>	<b>17</b>
<b>3.1.1 Criteria that depend only on each patient's outcome: .....</b>	<b>17</b>
<b>3.1.2 Criteria that depend only on each trial's outcome: .....</b>	<b>18</b>
<b>3.1.3 Criteria that depend on each trial's outcome compared to the "truth": .</b>	<b>18</b>
<b>3.2 PROPOSED EVALUATION CRITERIA BASED ON UTILITY FUNCTIONS .....</b>	<b>20</b>

3.2.1	Expected total personal utility.....	20
3.2.2	Expected societal utility.....	24
3.2.3	Expected total utility .....	28
4.0	<b>CTDESIGNEXPLORER – AN ACTION QUEUE-BASED OPEN-SOURCE SIMULATION EXPERIMENT PLATFORM FOR EVALUATING CLINICAL TRIAL DESIGNS.....</b>	<b>30</b>
4.1	<b>INTRODUCTION .....</b>	<b>30</b>
4.2	<b>CTDESIGNEXPLORER .....</b>	<b>32</b>
4.2.1	Overview.....	32
4.2.2	Simulation framework .....	34
4.2.3	Interoperability.....	36
4.2.4	Implementation: S4 classes and methods .....	38
4.2.5	Regular use: using existing classes and methods .....	41
4.2.6	Advanced use: extending, reusing and sharing .....	45
4.2.6.1	Extending .....	45
4.2.6.2	Reusing.....	47
4.2.6.3	Sharing .....	47
4.2.7	Code Validation .....	48
4.2.7.1	Code validation for “A+B” design.....	49
4.2.7.2	Code validation for the CRM.....	52
4.3	<b>FUTURE DEVELOPMENT.....</b>	<b>56</b>
5.0	<b>SOFTWARE APPLICATION IN EVALUATING EARLY PHASE CLINICAL TRIAL DESIGNS.....</b>	<b>57</b>

<b>5.1</b>	<b>LOGIT MODEL IN THE CONTINUAL REASSESSMENT METHOD....</b>	<b>57</b>
5.1.1	The continual reassessment method (CRM) .....	57
5.1.2	Re-parameterized logit model .....	58
5.1.3	Choices of prior distribution for $p_s$ .....	68
5.1.3.1	Experiment set-up .....	68
5.1.3.2	Results .....	73
5.1.3.3	Conclusions .....	85
5.1.4	Choices of prior distribution for $\gamma$ .....	86
5.1.4.1	Experiment set-up .....	86
5.1.4.2	Results .....	90
<b>5.2</b>	<b>PATIENT HETEROGENEITY .....</b>	<b>97</b>
5.2.1	The CRM .....	102
5.2.2	The standard “3+3” design .....	104
5.2.3	Conclusions .....	105
<b>6.0</b>	<b>CONCLUSIONS AND FUTURE WORK .....</b>	<b>106</b>
	<b>BIBLIOGRAPHY .....</b>	<b>108</b>



## LIST OF TABLES

Table 3.1 Four Types of Utility Functions for the Bivariate Outcome.....	21
Table 3.2 A Two-way Table for Different Kinds of Expected Utilities .....	29
Table 4.1 Examples of Common Names for Data Elements .....	37
Table 4.2 The Outcome Model Used in the Code Validation for “A+B” Design .....	50
Table 4.3 First Comparison Result in the Code Validation for “A+B” Design.....	51
Table 4.4 Second Comparison Result in the Code Validation for “A+B” Design .....	51
Table 4.5 Third Comparison Result in the Code Validation for “A+B” Design .....	51
Table 4.6 Designs Used in the in the Code Validation for the CRM.....	52
Table 4.7 First Comparison Result in the Code Validation for the CRM.....	54
Table 4.8 Second Comparison Result in the Code Validation for the CRM .....	55
Table 4.9 Third Comparison Result in the Code Validation for the CRM.....	55
Table 4.10 Fourth Comparison Result in the Code Validation for the CRM .....	55
Table 5.1 Different Choices of Values/Prior Distributions for $\alpha$ .....	63
Table 5.2 Initially Guessed Toxicity Probabilities .....	70
Table 5.3 CRM Designs with Different Priors for $p_s$ .....	70
Table 5.4 Toxicity Probabilities from the Dose-toxicity Curves with $\gamma=1$ .....	71
Table 5.5 Number of Toxicity Responses when “Guess=Truth” .....	78

Table 5.6 Number of Toxicity Responses when “Guess<Truth” .....	78
Table 5.7 Number of Toxicity Responses when “Guess>Truth” .....	79
Table 5.8 Proportion of Toxicity Responses when “Guess=Truth” .....	80
Table 5.9 Proportion of Toxicity Responses when “Guess<Truth” .....	80
Table 5.10 Proportion of Toxicity Responses when “Guess>Truth” .....	80
Table 5.11 Percentage of Patients Treated when “Guess=Truth” .....	81
Table 5.12 Percentage of Patients Treated when “Guess<Truth” .....	81
Table 5.13 Percentage of Patients Treated when “Guess>Truth” .....	82
Table 5.14 Proportion of Correct Estimation of the True Target when “Guess=Truth” .....	82
Table 5.15 Proportion of Correct Estimation of the True Target when “Guess<Truth” .....	83
Table 5.16 Proportion of Correct Estimation of the True Target when “Guess>Truth” .....	83
Table 5.17 Proportion of Early Stopping when “Guess=Truth” .....	84
Table 5.18 Proportion of Early Stopping when “Guess<Truth” .....	85
Table 5.19 Proportion of Early Stopping when “Guess>Truth” .....	85
Table 5.20 CRM Designs with Different Priors for $\gamma$ .....	87
Table 5.21 Toxicity Probabilities from the Dose-toxicity curves with Different $\gamma$ .....	88
Table 5.22 Number of Toxicity Responses when “ $\gamma=1$ ” .....	91
Table 5.23 Number of Toxicity Responses when “ $\gamma=2$ ” .....	92
Table 5.24 Number of Toxicity Responses when “ $\gamma=0.5$ ” .....	92
Table 5.25 Proportion of Toxicity Responses when “ $\gamma=1$ ” .....	93
Table 5.26 Proportion of Toxicity Responses when “ $\gamma=2$ ” .....	93
Table 5.27 Proportion of Toxicity Responses when “ $\gamma=0.5$ ” .....	93
Table 5.28 Percentage of Patients Treated when “ $\gamma=1$ ” .....	94

Table 5.29 Percentage of Patients Treated when “ $\gamma = 2$ ” .....	94
Table 5.30 Percentage of Patients Treated when “ $\gamma = 0.5$ ” .....	94
Table 5.31 Proportion of Correct Estimation of the True Target when “ $\gamma = 1$ ” .....	95
Table 5.32 Proportion of Correct Estimation of the True Target when “ $\gamma = 2$ ” .....	95
Table 5.33 Proportion of Correct Estimation of the True Target when “ $\gamma = 0.5$ ” .....	95
Table 5.34 Proportion of Early Stopping when “ $\gamma = 1$ ” .....	96
Table 5.35 Proportion of Early Stopping when “ $\gamma = 2$ ” .....	96
Table 5.36 Proportion of Early Stopping when “ $\gamma = 0.5$ ” .....	97
Table 5.37 Schemes in Consideration.....	101
Table 5.38 Different Maximum Sample Sizes in the CRM.....	103
Table 5.39 Expected Total Personal Utility Using the CRM.....	103
Table 5.40 Expected Societal Utility Using the CRM.....	103
Table 5.41 Expected Total Personal Utility Using the “3+3” Design .....	104
Table 5.42 Expected Societal Utility Using the “3+3” Design.....	104

## LIST OF FIGURES

Figure 3.1 Dependence Diagram for the Total Personal Utility .....	22
Figure 3.2 Graphical Display for a Societal Utility Function.....	25
Figure 3.3 Dependence Diagram for the Societal Utility .....	26
Figure 4.1 Overview of CTDesignExplorer.....	33
Figure 4.2 Action Queue-based Simulation Framework .....	35
Figure 4.3 Communications among Objects.....	36
Figure 4.4 CT Data Structure.....	40
Figure 4.5 Partial Documentation for Class "APlusBSpecifier".....	44
Figure 5.1 The relationship of $p_s$ and $\gamma$ to the Assumed Dose-toxicity Curve .....	62
Figure 5.2 Prior Density Plot of $p_s$ when $\alpha$ is Fixed.....	64
Figure 5.3 Enlarged Prior Density Plot of $p_s$ when $\alpha$ is Fixed .....	64
Figure 5.4 ECDF Plot of $p_s$ when $\alpha$ is Fixed .....	66
Figure 5.5 Prior Density Plot of $p_s$ when $\alpha$ is Free.....	66
Figure 5.6 ECDF Plot of $p_s$ when $\alpha$ is Free.....	67
Figure 5.7 Initially Guessed Dose-toxicity Response Curve .....	69
Figure 5.8 Prior Density Plots of $p_s$ .....	71

Figure 5.9 Underlying Dose-toxicity Curves with $\gamma=1$ .....	72
Figure 5.10 Dose Assignments for the First Five Patients in the “Informative” Design.....	74
Figure 5.11 Dose Assignments for the First Five Patients in the “Mild” Design.....	75
Figure 5.12 Dose Assignments for the First Five Patients in the “Bimodal” Design.....	75
Figure 5.13 Dose Assignments of the First Five Patients in the “Uniform” Design .....	76
Figure 5.14 Dose Assignments for the First Five Patients in the “Bimodal, restriction” Design.	77
Figure 5.15 Prior Density Plots of $\gamma$ .....	88
Figure 5.16 Underlying Dose-toxicity Curves with Different $\gamma$ .....	89
Figure 5.17 Dose Assignments for the First Five Patients in the “Small variance” or “Fixed” Design.....	90
Figure 5.18 Dose Assignments for the First Five Patients in the “Big variance” Design .....	91
Figure 5.19 Contour Plot for Dose Thresholds Distribution.....	99
Figure 5.20 Dose-Toxicity Curve .....	100
Figure 5.21 Dose-Efficacy Curve .....	100

## **ACKNOWLEDGEMENTS**

I am very grateful for all the learning opportunities provided to me, including delving into research, engaging in real data analyses, contending with homework, and observing the task approaches of others.

I would like to thank my advisor, Dr. Roger Day, for his guidance, encouragement and continuous support throughout the course of this dissertation. He is a very easy-going and helpful advisor. I enjoy a lot working with him. It was he who led me to the area of statistical software development, and it is also he who encourages me to think bravely and creatively. He always regards his students' needs as top priorities. I appreciate very much that he has sacrificed part of his vacation in order to read my dissertation draft and to give me timely comments.

Special thanks go to my supervisor as well as my committee member, Dr. Daniel Normolle, for his financial support and guidance. He has kindly supported my dissertation work since February, 2010 through one of his NCI R01 grants. With his financial support, I can stay more focused on my dissertation. He is the one who introduced me the logit model problem in the CRM design. The investigation of this problem has become an important part of my dissertation.

I would also like to thank my other two committee members, Dr. Hussein Tawbi and Dr. Abdus Wahed, for their valuable advice and generous accessibility during the course of this dissertation.

Last, but not least, I thank my family and friends for their love and support. Without them, I could not have finished this dissertation.

## **1.0 INTRODUCTION**

### **1.1 STATEMENT OF PROBLEMS**

In the past few decades, clinical trial (CT) designers have proposed many designs for trials at different stages of drug development, ranging from preclinical to Phase IV trials. This abundance of designs mandates a question for the investigators and statisticians planning a trial: what design would be the “best” for their trial? Determining the answer must begin with careful consideration of the criterion by which to judge designs, which should reflect the goals of the trial.

CT simulation is used by pharmaceutical companies and FDA to improve the efficiency and accuracy of drug development [1-4]. Sophisticated commercial software for trial simulations is available for those with resources to cover fees and with design challenges matching the software’s capabilities. The source code of commercial software is proprietary, so users have to believe the software does what it claims. Academic research centers usually use locally developed software mainly due to cost and flexibility considerations. Cost issues are obvious. For example, the recent quote price of a one-year single academic license for Pharsight trial simulator is \$11,235. Flexibility is needed primarily to explore novel designs and novel evaluation criteria. This local software development focuses on answering specific research questions in compressed time frames, and is not routinely sharable.



Open source software (OSS) is computer software for which the source code and certain other rights normally reserved for copyright holders are provided under a software license that meets the Open Source Definition [5] or that is in the public domain. OSS development approach has helped produce reliable, high quality software quickly and inexpensively. Besides, it offers the potential for a more flexible technology and quicker innovation. OSS covers a myriad of uses - from enterprise ecommerce to academic research, for example, Linux, Apache, Firefox and R packages. Open source solution may disseminate the innovative ideas in CT design efficiently.

When evaluating designs for a particular clinical trial, CT designers usually have some prior information from previous studies on the sampling distribution for patient characteristics which are relevant to the patient outcomes and/or patient-level decision-making in a clinical trial. There may be some knowledge or belief about the biological mechanisms by which outcomes arise, which may point to possible model types (e.g. logit or exponential models for the dose-toxicity response relationship), the probability of each model type, and the distribution for the model parameters. The classical way to utilize such prior information is to select several scenarios (models with specific parameter assignments), with the hope that they sufficiently represent the range of possible truths, and then to evaluate designs under each of these scenarios, applying requirements such as a cap on the type I and type II errors. This classical approach has problems: a small set of scenarios may not be sufficient, and the methods for synthesizing the evaluation results from different scenarios are somewhat artificial. The expected utility paradigm of Bayesian decision theory is well suited to helping CT designers with incorporating such prior information, and providing a comprehensive evaluation of design operating characteristics. However, an extensive literature review of existing design evaluation criteria did not reveal the criteria based on the expected utility.

Patients enrolled in Phase I cancer trials vary greatly in type of cancer, numbers and types of previous treatment, age, sex, genetic profile and many other factors that may impact their tolerance to the testing treatment. Ignoring patient heterogeneity in the Phase I trials may do harm to patients or recommend an either suboptimal or too toxic dose for future studies. In practice, however, few Phase I clinical trials account for patient heterogeneity. This dissertation explores the effects of patient heterogeneity on CT design performance.

The Phase I dose-finding designs using the continual reassessment method (CRM) present challenges and opportunities. It was first introduced by O’Quigley et al. in the year 1990 [6], and has inspired many variations [7-14]. The CRM and its variations assume the probability of toxicity response increase monotonically with dose via a parametric model. They typically apply Bayesian methods and assign prior distributions to the parameters in the model, although maximum likelihood methods have been proposed [15]. Single-parameter models are usually used and the choice of the intercept values seems arbitrary. There has been a divergent opinion about the number of parameters used in the model [16-20]. Supporters for single-parameter models argue that they can adequately approximate the dose-toxicity response relationship by a single parameter in the range of true target dose and that it is not possible to reliably estimate two or more parameters in Phase I clinical trials where sample sizes are small. However, as Phase I trials grow in complexity, for example, late-onset toxicity response and combined therapy, the single-parameter model may not be rich enough to describe underlying outcome model sufficiently. A major challenge is how to frame Bayesian priors for the one- and multi- parameter models. We find that the use of rescaled doses, to be discussed in Section 5.1.1, obscures the interpretations of parameters, making specification of priors unnecessarily difficult. Using easily interpretable parameters would help to choose the number of free parameters in a model, and to

set up sensible prior distributions that genuinely reflect the investigators' prior beliefs. Thus, efforts are necessary to define dose-toxicity models with more interpretable parameters in the CRM.

## **1.2 OBJECTIVES OF THE STUDY**

The primary objective of this dissertation was to build a transparent, extendible simulation experiment platform and to provide standards for further development so that CT designers can evaluate available designs and/or share their innovations. Five specific objectives are described briefly as follows:

**Objective 1:** Review Phase I dose-finding clinical trial designs.

**Objective 2:** Review the existing Phase I design evaluation criteria and present new criteria for evaluating the expected total personal utility, societal utility and total utility.

**Objective 3:** Build an action-queue based open-source extendible simulation experiment platform, and develop it into an R package.

**Objective 4:** Re-parameterize the logit model in the CRM for more interpretable parameters, and investigate the choices of prior distribution using the platform.

**Objective 5:** Quantitatively demonstrate the effect of ignoring patient heterogeneity on the CRM and the standard “3+3” design with respect to the expected total personal utility and societal utility.

## **2.0 OVERVIEW OF PHASE I DOSE-FINDING CLINICAL TRIAL DESIGNS**

Phase I trials are typically very small, uncontrolled and sequential studies of human subjects designed to determine the recommended dose of an experimental drug for the subsequent Phase II trials. Probably because of the non-randomization and small sample size, statistical considerations were ignored for many years in the Phase I trial designs. Nowadays, the statistical input is becoming more and more important in designing Phase I trials to more accurately and efficiently choose a dose for the subsequent Phase II trials while minimizing the patient risks. In addition to the toxicity risk, the patients in Phase I cancer trials also incur the risk of receiving sub-therapeutic doses, which is of less concern in Phase I trials for milder diseases. The goal of this chapter is to provide a general introduction to the Phase I trial and its designs, and to review the recent development of the Phase I dose-finding cancer trial designs between 2007 and 2010.

### **2.1 INTRODUCTION**

Phase I clinical trials are the first step in testing a new treatment in people. They can be first-in-human trials, or new studies of an agent or agents previously evaluated in humans, which includes new agent formulations, routes of administration, combinations of agents and etc. One primary objective of the Phase I trials is to determine an appropriate dose for the Phase II trials. That dose is usually called maximum tolerated dose (MTD) with the assumption that the

probabilities of having toxicity and/or efficacy responses increase with the increasing dose. This assumption is generally valid when the testing drugs are cytotoxic agents, however, it may fail in some biologic agents. More general terms, for example, Phase II recommended dose (P2RD), recommended Phase II dose (RP2D) and optimal dose (OD), have been used in the literature. In this dissertation, we use the term OD to refer to the true target dose that a particular Phase I trial should recommend for the subsequent Phase II trial under a specific criterion given complete knowledge, i.e. OD is the dose that optimizes a specific criterion if we know everything; we use RP2D as the dose that is recommended for the subsequent Phase II trial at the end of a particular Phase I trial. The second primary objective of the Phase I trials is to identify the toxicities associated with the testing treatment, particularly the dose-limiting toxicities (DLTs). DLT is usually defined as adverse events sufficiently morbid that constitute a practical limitation to the delivery of the treatment. In general, DLTs are grade  $\geq 3$  non-hematological and grade 4 hematological toxicities (the latter not including grade 3 because hematological toxicities have become generally easier to manage). The “Common Terminology Criteria for Adverse Events” (CTCAE, v3.0) are currently used to grade toxicities. The CTCAE are available at the website (<http://ctep.cancer.gov>) of the National Cancer Institute’s Cancer Therapy Evaluation Program (CTEP).

The designs for the Phase I trials fall into two categories. Algorithm-based designs, which do not assume any OD, and regard RP2D as a statistic calculated directly from data; Model-based designs, which have OD as a parameter, and regard RP2D as an estimate of the OD. Compared to the algorithm-based designs, model-based designs can improve the efficiency if the model assumptions are satisfied. However, when the model assumptions are incorrect, model-based designs can lead to an inaccurate estimate of OD, and it may either overdose or

underdose participants in the trials unnecessarily. Therefore, robustness of a model-based design to the model misspecifications is important to check. Some two-stage designs mix algorithm- and model- based methods, with the first-stage algorithm-based and the second model-based [5-7]. These mixture designs intend to begin the second stage near the OD, where model-based methods may work better.

Eisenhauer et al.'s review paper [21] declares three components to a Phase I trial design: (1) the starting dose, (2) the number of patients per dose level, and (3) the dose-escalation scheme. One-tenth of the mouse LD<sub>10</sub> (the dose that was lethal to 10% of animals) has historically been selected to be a safe starting dose in humans, as long as that dose is not lethal or life-threatening to a second species (e.g. rat, dog) [22]. Increasing the starting dose could potentially reduce the trial length and decrease the number of patients receiving ineffective doses. Attempts have been tried to find a higher safe starting dose [21], notably through the use of interspecies scaling [23]. Decreasing the number of patients per dose level would limit the number of patients exposed to very low doses and might shorten the trial length if the recruitment of patients per dose level is rate limiting. However, it could be problematic when there are only few patients per dose level (e.g. a single patient per dose level) in some trials where the patient population is very heterogeneous or the pharmacokinetic measurements are required. Finally, the dose-escalation scheme has attracted considerable attention from statisticians. More aggressive dose-escalation in the initial portion of the trial may shorten the length of a trial but may cause more patients to experience serious toxicity responses. Therefore, risks and benefits affected by the aggressiveness of dose-escalation need careful consideration [24].

Investigators conducting Phase I trials must adhere to the ethical norms of clinical research [25-31]. Participants in Phase I cancer trials are almost always patients with refractory disease or for whom there is no standard therapy, often at a very high risk death, who consent to participate in the trial only as a last resort in seeking a cure. These trials raise special ethical issues in the dose- escalation process because of a fundamental conflict: on the one hand, there is a need to avoid a large jump from a dose with no observable toxicity to a lethal dose; on the other hand, there is a need to go rapidly, so that large numbers of patients are not treated at ineffective doses. One of the attempts to resolve this conflict is to construct a utility function associated with the bivariate outcome (toxicity and efficacy responses) [32] and select the dose that maximizes the expected utility for each patient.

A number of Phase I trial designs have been proposed in the past few decades ([6, 8, 9, 15, 33-36], among others). Despite a number of criticisms against the standard 3+3 design, in practice, it is still the most frequently used probably because it is easy to understand and implement. More recent designs often target specialized situations, such as late-onset toxicity, combined agents, patient population heterogeneity and ordinal toxicity outcomes. Several authors have provided comprehensive reviews of the Phase I trial designs prior to the year 2007 [21, 22, 24, 37-48]. The purpose of this chapter is to review the Phase I dose-finding cancer trial designs which were proposed between the year 2007 and 2010. Section 2.2.1 reviews the designs which consider only a single outcome (usually toxicity) and section 2.2.2 reviews designs considering a bivariate outcome (toxicity, efficacy). We review papers in the sense of original terminology and notation, which can be variable across papers.

## 2.2 RECENT DEVELOPMENT

### 2.2.1 Designs for single outcome

Most Phase I dose-finding designs dichotomize toxicity grades based on DLT, wherein the outcome is regarded as “toxicity” if a patient experiences DLT, and “non-toxicity” if a patient does not experience DLT. However, this dichotomization may discard a lot of useful information [41] , for example, by equating life-threatening, irreversible, or long-duration toxicities with others[49]. Liu et al. [50] presented a parametric sequential design based on the proportional odds model with ordinal toxicity response in a discrete dose space. A simple and flexible penalty function was used in the cumulative information matrix to construct a penalized local optimality criterion for finding multiple quantiles of ordinal data. Ivanova and Kim [51] proposed a unified approach to dose finding in the studies where the quantity of interest is a monotone function of the dose and the goal is to estimate the dose corresponding to a pre-specified desired value of the function. The function may be a certain weighted sum of rates of various toxicity grades or the expected value of a continuous outcome. In their design, dose-escalation depends on the normalized difference between the current dose and the target dose at which the outcome of interest is equal to the pre-specified value. Yuan et al. [49] described a simple way to incorporate multiple ordinal toxicity grades using quasi-Bernoulli likelihood and then couple this likelihood with the continual reassessment method (CRM) to dose finding. They measured the relative severity of different toxicity grades by equivalent toxicity (ET) score, normalized the ET scores against the maximum and obtained the quasi-maximum likelihood estimate (QMLE) of MTD. Potthoff and George [52] treated toxicity response as a continuous variable, ranging from 0 to 1. They defined MTD as the dose with the targeted mean toxicity. The distribution of toxicity given



dose was assumed to follow a beta distribution with mean modeled by a two-parameter logistic model. Their design chooses the dose for each patient that maximizes the weighted sum of two quantities,  $(1 - \text{posterior toxicity estimate})$  and the inverse of the posterior estimated variance of MTD.

Many modified or extended CRM designs have been proposed ever since the original CRM proposal appeared in the year 1990 [6]. A one-parameter power model,  $\{(p_1^\alpha, \dots, p_J^\alpha) : \alpha \in (0, \infty)\}$  is usually used in CRM for modeling toxicity probabilities, and clinicians need to pre-specify the values  $\{p_j, j = 1, \dots, J\}$ . The choice of  $\{p_j, j = 1, \dots, J\}$  affects design properties. An unwillingness to pre-specify  $\{p_j, j = 1, \dots, J\}$  is one of the obstacles for clinicians to accept the CRM method. To overcome this arbitrariness and further enhance the robustness of the CRM design, Yin and Yuan [53] suggested using multiple parallel power models with different pre-specifications of  $\{p_j, j = 1, \dots, J\}$  and possibly different prior distribution of  $\alpha$ . A Bayesian model averaging (BMA) approach is then applied to estimate the posterior probabilities of toxicity. CRM has been criticized for exposing too many patients to doses higher than the MTD if the model or prior is not correctly specified. Babb et al. [54] proposed the escalation with overdose control (EWOC) method to minimize the number of overdosed patients. However, their dose escalation is slower and more patients receive sub-therapeutic doses compared to CRM. Chu et al. [55] unified CRM and EWOC approaches and presented a hybrid design to combine the strengths from both approaches. The original CRM described by O’Quigley [6] used single-patient cohorts which needs fewer patients but may take longer time. CRM has been extended to allow for multiple-patient cohorts (usually 3 patients) to shorten the trial duration but may need more patients or reduce the accuracy of the MTD estimate [8]. Huang and Chappell [56] introduced three-dose-cohort designs which randomly

administer three patients in the same cohort with three different doses (low, medium and high) according to the quantiles of the posterior distribution of the parameter in dose-response model. Their simulation results demonstrated their designs combined some advantages from both single-patient cohort CRM and three-patient cohort CRM where the patients in the same cohort receive the same dose.

Time-to-Event CRM (TITE-CRM) [14] is a popular method to address the problem of the late-onset toxicities. Bekele et al. [57] modified TITE-CRM by providing rules for suspending accrual if the predicted risk of toxicity is unacceptably high. This modification improves the safety of the trial at the price of longer duration on average.

When the toxicity is reversible and the carry-over effect of doses can be diminished after a short washout period, intra-patient dose escalation can be useful for more effectively providing information on inter-patient variability; it also increases the chances that patients receive therapeutic doses. Fan and Wang [58] proposed an “m-dose design” which treats patients having no toxicity response at their current doses with the next higher dose until they have received m doses or their current dose is the highest dose. They also extended Leung and Wang’s method [59] from the single-dose to the “m-dose design” and obtained an iterative nonparametric estimator for the probabilities of toxicities.

Dose-finding in the Phase I trials is usually regarded as an estimation problem. In some Phase I trial designs, the precision (inverse variance) of the estimated OD plays the role of utility function or the condition for the stopping rule. Most of the designs assume a monotonic increasing dose-toxicity relationship. The design properties worsen seriously if this assumption is not valid. Cheung [60] explored a class of algorithm-based designs based on two-stage stepwise testing procedures without the monotonicity assumption. The goal of this design is to identify

the MTD with high probability while keeping the probability of choosing an unsafe dose low. The stepwise test is defined with respect to the family of hypotheses,  $H_{0i} : p_i \geq \phi$  versus  $H_{1i} : p_i < \phi$  for  $i = 1, \dots, K$ , where  $p_i$  is the true toxicity probability at the  $i$ th dose level and  $\phi$  is some toxicity probability which is a little higher than the target toxicity rate but still within the safe range. Their defined family-wise error rate (FWER) and PCS are used to control the type I and type II errors of this stepwise testing procedure. Extensive simulations demonstrate their designs have competitive operating characteristics under a wide range of scenarios but may require larger than typical sample size.

Zandvliet et al. [61] proposed a two-stage model-based design for finding the recommended Phase II doses for various regimens of the same drug. The first stage of their design is the conduct of a Phase I study for a single regimen using conventional modified Fibonacci-like dose escalation. Then a pharmacokinetic and pharmacodynamic population model is developed from the data collected in the first stage, and used for performing a simulation study. They use the five percentile of recommended doses obtained from the simulation study as the starting dose for the Phase I studies with other regimens. After the conduct of all the Phase I studies, the recommended doses for further studies are determined through a post-hoc analysis. They demonstrated via simulation that their design may help to reduce the number of patients treated under sub-optimal doses, and to increase the precision of dose selection for Phase II evaluation compared to the conventional design.

To shorten the study duration, Skolnik et al. [62] proposed a “rolling six” design which allows for accrual of two to six patients concurrently onto a dose level based on the number of currently enrolled evaluable patients and the number of DLTs. Simulations shows that their

design considerably decreases the trial duration without increasing the toxicity risk and sample size significantly compared to the standard 3+3 design.

Combined therapy has long been a commonplace in cancer clinical trials due to its occasional dramatic successes, together with hopes of potential synergistic therapeutic effects and non-overlapping toxicities. The toxicity ordering in the multi-dimensional dose combination space is typically unknown. Thus, imposing the usual assumption of non-decreasing dose-response relationship in the Phase I trials with combined agents requires adopting some ordering with little or no justification. This poses a big design challenge for biostatisticians to efficiently, accurately and ethically choose the appropriate dose combination for the subsequent Phase II trial. A popular dose-search path is to first escalate one agent while the other agents are at the lowest dose level, then escalate the second agent while the first agent is at the highest acceptable dose level and the other agents are at the lowest dose level, and keep doing until the last agent is tested. Even though this approach is easy to understand, it may miss the best combination of the agents. Fan et al. [63] presented a strategy for searching MTD (here, D represents dose combination) within all possible dose combinations in a two-agent combination clinical trial. They provided two versions of design. One is a two-stage design in which the first stage selects MTD candidates and the second stage identifies an MTD among them; The other, more complex, has a preliminary stage to find a starting dose combination, along the diagonal of the tier dose plane, based on the observations of toxicity responses at and below DLT. Subsequently, escalation can only move up one dose level of one drug at each step. Yin and Yuan published two papers [64, 65] in the year 2009 for combined agents dose-finding trial designs. Like CRM, they assumed that the toxicity rate at the  $j$ th dose level for the  $i$ th agent is in the form of  $p_{ij}^{\alpha}$ , where the  $p_{ij}$  are physician-specified initial guesses for the toxicity rates and  $\alpha$  is an unknown

positive parameter to enhance the flexibility and to accommodate physicians' uncertainty. Dose escalation or de-escalation is restricted to changing one agent by no more than one dose level. The Bayesian dose-finding algorithm is based on the fixed probability cut-offs for dose escalation and de-escalation. They proposed [64] a Bayesian adaptive design which models the binary toxicity outcomes through a series of 2x2 contingency tables. They jointly model the probabilities in the contingency table at each dose combination using the Gumbel model [66], where the association parameter characterizes the agent synergistic effect. They also proposed [65] a different Bayesian adaptive design for dose-finding trials combining more than two agents. The synergistic effect of combined agents is modeled using a copula-type regression.

### **2.2.2 Designs for bivariate outcome**

Patients in Phase I cancer trials are very heterogeneous in the disease type and extent. However, most of the designs proposed or used in the real Phase I clinical trials ignore this inter-subject variability. Thall et al. [36] extended the Thall and Cook method [67] by accounting for patient covariates and dose-patient covariate interactions. They used historical data to obtain an informative prior on the patient covariate main effects and assumed uninformative priors for all dose effect parameters. Their method is very complex and computationally intensive. It will require a substantial effort from both statistician and physicians planning the trials. Nonetheless, their simulation results show that ignoring the patient heterogeneity in the dose-finding trials runs a high risk of assigning inferior doses to particular patient subgroups. Further research to develop simpler designs accounting for patient heterogeneity is warranted.

Ivanova et al. [68] presented an adaptive dose-finding strategy based on both toxicity and efficacy in a crossover study setting. In their example, efficacy and toxicity are measured by a

continuous variable and binary variable respectively. They assume that the mean efficacy response increases with doses in the lower dose range and may possibly decrease with doses in the higher dose range, while toxicity rate is non-decreasing throughout the dose range. They define a utility function which takes into account both efficacy and toxicity responses. Their design concentrates on assigning more patients at and around the dose which maximizes the utility function. Application of their design to a Merck study showed that their proposed design saves sample size compared to the traditional design (equal allocation crossover design) while maintaining similar power for the primary comparison with placebo.

Dragalin et al. [69] introduced a two-stage design based on the bivariate probit dose-response model, using optimal experimental design methodology to construct a dose allocation procedure balancing efficiency with the desire to maximize the number of patients receiving both safe and efficacious doses. The efficiency is inversely proportional to the variance of the estimator for the target dose. They presented some reasonable penalty functions to add to the optimality criterion to address concerns with respect to the ethics and cost. This two-stage design was shown to perform fairly well under any initial design in the first stage.

Mandrekar et al. [70] presented an adaptive design for two-agent Phase I trials based on an extended continuation ratio model from the single-agent “TriCRM design” proposed by Zhang et al. [71]. Three mutually exclusive and exhaustive outcomes are considered in their design: no response (no efficacy and no toxicity), success (efficacy and no toxicity) and toxicity. Note that this classification of outcomes regards response as irrelevant if toxicity occurs. Joint criteria based on both toxicity and success determine the dose combination for the next cohort. Huang et al. [72] introduced a novel parallel Phase I/II trial design for combined agents. There are three new aspects in their design: first, the trial begins with an initial dose-escalation period

based on the number of DLTs to select admissible dose combinations for testing; secondly, they randomize patients adaptively to all the admissible dose combinations with the assignment probabilities proportional to the credibility of the supposition that one dose combination is superior to the other ones; and thirdly, they perform interim monitoring for efficacy and toxicity to temporarily close dose combinations with lower efficacy and eliminate those with intolerable toxicity. The simulations show that their proposed design has better efficiency compared to the conventional designs.

Wang and Day [73] presented a new adaptive Bayesian Phase I design using a joint dose thresholds model. Their design utilizes prior information about the joint toxicity and efficacy dose thresholds distribution as well as accumulating data. The dose chosen for each patient maximizes the current posterior expected personal utility. They provide four types of utility functions based on a patient's bivariate outcome. Simulation shows that their design, compared to the standard design, identifies the right dose with smaller sample size and has more patients experiencing the desirable outcome (non-toxicity and efficacy responses) and fewer patients experiencing the undertreated outcome (non-toxicity and non-efficacy responses).

### **2.2.3 Summary**

In this review, we have seen that the designs that people propose differ in many ways. Among the most important is the degree to which they use available prior and ongoing information to choose each patient's dose. The majority of the recently proposed designs continue to incorporate the Bayesian calculations in various ways. Increasingly the designs include an initial algorithm-based dose-escalation period when there is little prior information the drug under study.

### **3.0 EVALUATION CRITERIA FOR PHASE I DOSE-FINDING CLINICAL TRIAL DESIGNS**

Evaluation of CT designs is a very important procedure when planning a clinical trial, modifying trial designs in response to interim results, or comparing competing designs in methodology research. The choice of criteria for evaluation depends on the goal of the studies. For example, if the goal of a certain Phase I trial is to find OD as accurately as possible, one evaluation criterion might be bias of the RP2D. Since Phase I trials are usually first-in-human studies, evaluating operating characteristics of designs before running clinical trials is critically necessary from both ethical and scientific considerations. In this chapter, I review the existing evaluation criteria for Phase I dose-finding trial designs and introduce new criteria based on utility functions.

#### **3.1 EXISTING EVALUATION CRITERIA**

This section classifies evaluation criteria according to the inputs.

##### **3.1.1 Criteria that depend only on each patient's outcome:**

- Number and/or proportion of toxicity responses
- Number and/or proportion of efficacy responses



- Number and/or proportion of joint responses of toxicity and efficacy

### **3.1.2 Criteria that depend only on each trial's outcome:**

- Total number of included patients
- Number and/or proportion of patients treated at each dose level
- Number of cohorts to complete the trial. (The number of cohorts suggests the length of time to complete the trial when there are patients readily available for entry into the trial [74].)
- Probability of a dose being chosen as the RP2D
- Number and/or proportion of trials where no RP2Ds are determined
- Number and/or proportion of patients treated at the OD given a pre-specified criterion
- Number and/or proportion of patients treated below the OD given a pre-specified criterion
- Number and/or proportion of patients treated above the OD given a pre-specified criterion
- Variance of the RP2Ds
- Mean probability of having a toxicity response at the RP2D (It is called by Lin and Shih [75] target toxicity level) for algorithm-based designs.
- Variance of the estimated probability of having a toxicity response at each dose level

### **3.1.3 Criteria that depend on each trial's outcome compared to the “truth”:**

The “truth” can be:

- True dose-toxicity response curve
- OD given a pre-specified criterion
- The true probability or odds of toxicity responses at the RP2D
- Some optimal design used as gold standard

Criteria that are interpretable as expected loss:

- Mean absolute error of the RP2D compared to the OD
- Mean absolute error of the true probability of toxicity responses at the RP2D compared to that at the OD
- Proportion of the determined RP2Ds that are too low ( for example, Storer [76] considered the determined RP2D too low when its odds of toxicity responses is less than half that of the OD)
- Proportion of the determined RP2Ds that are too high ( for example, Storer [76] considered the determined RP2D too high when its odds of toxicity responses is more than twice that of the OD)
- Proportion of the inappropriately early terminated trials ( inappropriate early termination refers to the scenario that the trial has to be terminated because of the incorrect conclusion that the lowest testing dose is too toxic according to a pre-specified criterion)

Curiously, to my knowledge, mean squared error hasn't been used as a criterion in the Phase I design literature, even though it has an interpretation as an expected loss.

Criteria that are not interpretable as expected loss:

- Proportion of the trials where the OD are estimated correctly
- Bias of the determined RP2D compared to the OD given a pre-specified criterion

- Bias of the estimated probability of toxicity responses compared to the true probability of toxicity responses at each dose level
- Efficiency of the design compared to another design considered optimal or ideal proposed by Paoletti et al. [77].

### 3.2 PROPOSED EVALUATION CRITERIA BASED ON UTILITY FUNCTIONS

The above review of existing evaluation criteria did not reveal the use of evaluation criteria based on utility functions. Here, we propose flexible criteria for evaluating Phase I trial designs by assessing through CT simulation the expected total personal utility, societal utility and total utility. We acknowledge the challenge of developing a defensible and reasonable utility function that accounts for all the relevant risks and benefits.

#### 3.2.1 Expected total personal utility

We define total personal utility as the total individual net benefit (less harm) experienced by patients on a particular clinical trial. We propose to measure the total personal utility by the sum of utilities associated with each participating patient's outcome. Let  $U_p$  stand for the total personal utility. Suppose that the total number of patients on a particular trial is  $n$ , their outcomes are  $y_1, \dots, y_n$ , and the corresponding utilities associated with these outcomes are:  $u(y_1), \dots, u(y_n)$ .  $U_p$  is then calculated by:

$$U_p = \sum_{i=1}^n u(y_i) \quad (1)$$

For illustration, let us assume each patient has four possible outcomes:  $TE$  (toxicity and efficacy responses),  $Te$  (toxicity and non-efficacy responses),  $tE$  (non-toxicity and efficacy responses) and  $te$  (non-toxicity and non-efficacy responses). The corresponding utilities associated with these outcomes are  $u(TE), u(Te), u(tE), u(te)$  respectively. Table 3.1 lists four types of utility functions of the outcomes. The simple utility function only considers the benefit obtained from the best outcome  $tE$ , and ignores the difference among the other three possible outcomes. The additive utility function assumes that the utilities of having a toxicity, non-toxicity, efficacy, or non-efficacy response are -0.5, 0.5, 0.5, -0.5 respectively and that the utility of having joint responses (e.g.  $TE$ ) is the sum of the utilities of its respective responses. The aggressive utility function regards toxicity and non-toxicity responses as identical when a patient has an efficacy response, and it gives a penalty to having toxicity response only when a patient does not have an efficacy response. Whenever a patient has a toxicity response, the cautious utility function assigns the utility of -1 to the outcome. Note that all four types of utility function assign the utility of 1 to the best outcome ( $tE$ ), and 0 to the outcome  $te$ .

Table 3.1 Four Types of Utility Functions for the Bivariate Outcome

Outcome	$TE$	$Te$	$tE$	$te$
Simple Utility	0	0	1	0
Additive Utility	0	-1	1	0
Aggressive Utility	1	-1	1	0
Cautious Utility	-1	-1	1	0

Let  $d \in \mathcal{D}$  denote a particular design,  $m \in \mathcal{M}$  denote a particular interoperable model family pair (population and outcome model families) where “interoperable” means that a population model family can provide patient characteristics that an outcome model family

requires,  $\lambda_m \in \Lambda_m$  denote a vector of specific parameter values for model family pair  $m$ , and  $x \in \mathcal{X}$  refer to the observations (all patients' outcomes,  $\{y_i, i=1, \dots, n\}$ ), in a particular clinical trial.  $U_p$  is a function of  $x$ , and  $x$  depends on the design, underlying outcome generation mechanism and underlying patient characteristics (relevant to the design and outcome generation mechanism). Therefore  $U_p$  also depends on these three factors. Figure 3.1 displays the simple dependence relationship among those quantities, where each arrow points to the dependent quantity.

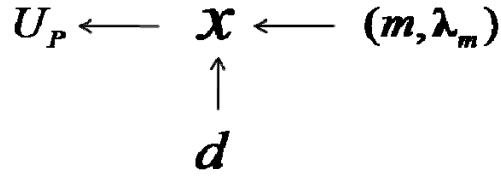


Figure 3.1 Dependence Diagram for the Total Personal Utility

For simplicity, we assume observations and parameters are all continuous. Let  $\pi_\lambda(\lambda_m)$  denote the prior density function for  $\lambda_m$ , and  $\pi_m(m)$  denote the prior probability of model family  $m$ . The expected total personal utility for a design  $d$  is:

$$\begin{aligned}
 EU_p(d) &= \sum_{m \in \mathcal{M}} \pi_m(m) \int_{\Lambda_m} \left\{ \int_{\mathcal{X}} U_p(x) f(x|d, m, \lambda_m) dx \right\} \pi_\lambda(\lambda_m) d\lambda_m \\
 &= \sum_{m \in \mathcal{M}} \pi_m(m) \int_{\Lambda_m} EU_p(d, m, \lambda_m) \pi_\lambda(\lambda_m) d\lambda_m \\
 &= \sum_{m \in \mathcal{M}} \pi_m(m) EU_p(d, m)
 \end{aligned} \tag{2}$$

, where (in an abuse of notation)  $EU_p(d, m, \lambda_m) = \int_{\mathcal{X}} U_p(x) f(x | d, m, \lambda_m) dx$  and

$$EU_p(d, m) = \int_{\Lambda_m} EU_p(d, m, \lambda_m) \pi_\lambda(\lambda_m) d\lambda_m .$$

We can follow the steps below to approximate  $EU_p(d)$  through the Monte Carlo simulation:

For a design  $d$  and a model family pair  $m$ , do steps 1 to 6:

1. Sample  $n_\lambda$  sets of parameters from the prior distribution  $\pi_\lambda(\lambda_m)$ :

$$\{\tilde{\lambda}_m^w, w = 1, \dots, n_\lambda, \tilde{\lambda}_m^w \in \Lambda_m\}$$

2. Simulate a clinical trial under  $d, m, \tilde{\lambda}_m^w$ , and get observations  $x_{dmw}$

3. Calculate the  $U_p(x_{dmw})$  using Equation (1)

4. Repeat steps 2-3  $n_r$  times

5. Repeat steps 2-4 for the rest sets of sampled parameters

6. Calculate  $\bar{U}_p(d, m) = \frac{1}{n_\lambda} \sum_{w=1}^{n_\lambda} \left\{ \frac{1}{n_r} \sum_{r=1}^{n_r} U_{p,r}(x_{dmw}) \right\}$ , which is an approximation to

$$EU_p(d, m)$$

Repeat steps 1-6 for the other possible model family pairs, then calculate  $\bar{U}_p = \sum_{m \in \mathcal{M}} \pi_m(m) \bar{U}_p(d, m)$ , which is the approximate estimate of  $EU_p(d)$ .

When  $m$  and  $\lambda_m$  are known and fixed, the computation for the expected total personal utility would be greatly simplified and it can be approximated using Equation (3). An example will be given in Section 5.2.

$$EU_p(d | m, \lambda_m) \approx \frac{1}{n_r} \sum_{r=1}^{n_r} U_{p,r}(x(d) | m, \lambda_m) \quad (3)$$

### 3.2.2 Expected societal utility

We define societal utility as the total net benefit from knowledge gained by the society after running a particular clinical trial, and we denote it as  $U_s$ . The primary goal of a Phase I dose-finding trial is to find/recommend a dose of a drug for the subsequent Phase II trial, with the hope that the drug at this dose can have the maximal efficacy probability while having tolerable toxicity probability. We propose to measure  $U_s$  by the utility associated with the chosen RP2D. Assuming RP2D is chosen from one of the testing doses, let  $P_E^\dagger$  and  $P_T^\dagger$  denote the probability of efficacy and probability of toxicity at the RP2D respectively. The current standard treatment is the treatment that patients will typically receive if they are not enrolled in the Phase I clinical trials. A family of reasonable societal utility functions is a function of  $P_E^\dagger$  and  $P_T^\dagger$  and it assigns the utility of 0 to the current standard treatment. Figure 3.2 shows the graphical display for a possible societal utility function of this family. The red point with higher  $P_E^\dagger$  and  $P_T^\dagger$  corresponds to the probability pair from the current standard treatment. Another red point on the  $P_E^\dagger$  axis is elicited from physicians and is assumed to have the same societal utility as the current standard treatment. These two red points define the reference line (here, we assume the contour lines are linear). Other lines are drawn parallel to this reference line within the two-dimensional domain  $[0,1]^2$ . The points on the same line have the same societal utility, the ideal point (1, 0) has the largest societal utility, and the societal utility of the probability pair gets smaller as it moves farther away from the ideal point.

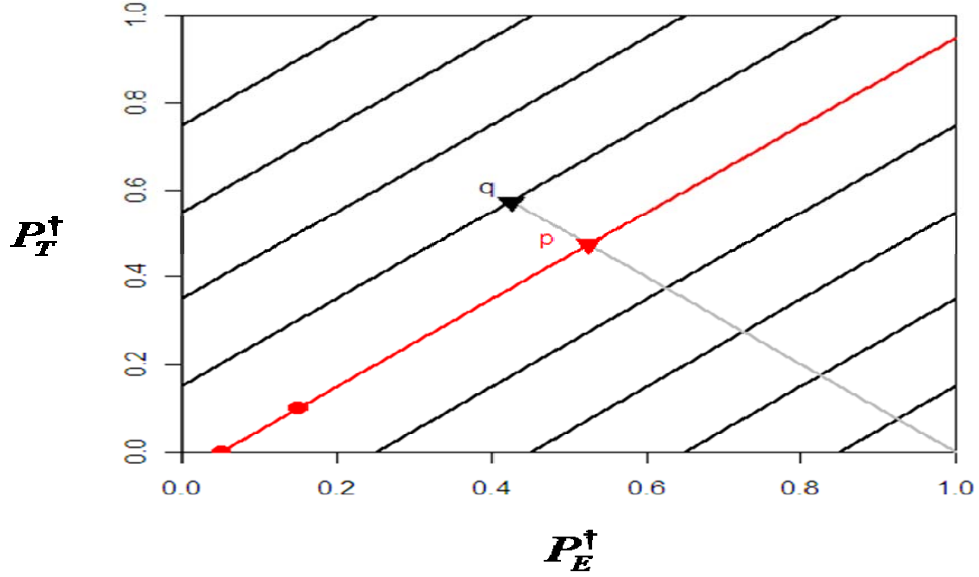


Figure 3.2 Graphical Display for a Societal Utility Function

Divergent opinions may arise with respect to the shape of the contour lines (linear or nonlinear) and the relationship between different contour lines (parallel or nonparallel). The investigation of that direction is out of the scope of this dissertation. Another tricky or challenging part in defining a societal utility function is the assignment of utilities to each contour line, which demands careful considerations of risks and benefits relative to the current standard treatment. Assuming the contour lines are parallel to each other, we propose a method to assign the utilities by adapting the method introduced by Thall and Cook [67]. Suppose  $q$  is a point on a contour line, and  $p$  is the point at which the line from the ideal point  $(1,0)$  to  $q$  crosses the reference line. The utility assigned to that contour line containing  $q$  is calculated by  $\rho(p) / \rho(q) - 1$ , where  $\rho(\cdot)$  denotes the Euclidean distance from a point in the contour line to the ideal point. The utility assigned to the reference line is 0. The contour lines closer to the ideal



point compared to the reference line have the positive utility, while the contour lines farther away from the ideal point have the negative utility.

The chosen  $RP2D$  from a Phase I clinical trial depends on the design and the observations, denoted by  $x^\bullet$ , which include all patients' outcomes ( $x$ ) and also their assigned doses.  $x^\bullet$  depends on the design, the underlying model family pair  $m$  (population and outcome model families) together with associated parameter values  $\lambda_m$ . Let  $\mathcal{Z} = \{z_j, j = 1, \dots, J\}$  denote the set of all testing doses, and  $\mathcal{Z}^\circ = \{NA_T, NA_I\}$  denote the set of possible missing values for  $RP2D$  in which  $NA_T$  refers to missing value because of too toxic lowest dose and  $NA_I$  refers to missing value because of still safe highest dose. If  $RP2D \in \mathcal{Z}$ , our proposed societal utility function is a function of  $P_E^\dagger$  and  $P_T^\dagger$ , and  $(P_E^\dagger, P_T^\dagger)$  depends on the chosen  $RP2D$ , underlying model family pair and its associated parameter values. Following the notation mentioned in Section 3.2.1, we draw Figure 3.3 which displays the dependence relationship among those quantities.

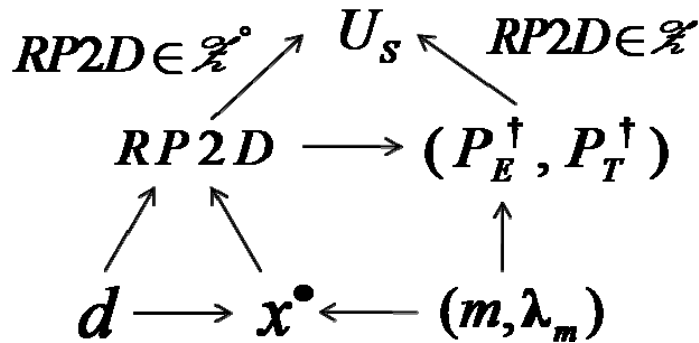


Figure 3.3 Dependence Diagram for the Societal Utility

For a design  $d$ , model family pair  $m$  and its parameter values  $\lambda_m$ , the expected societal utility  $EU_S(d, m, \lambda_m)$  is:

$$EU_S(d, m, \lambda_m) = \sum_{z \in \mathcal{Z}} \Pr(RP2D = z \mid d, m, \lambda_m) U_S(P_E^\dagger, P_T^\dagger \mid RP2D = z, m, \lambda_m) + \sum_{z \in \mathcal{Z}^c} \Pr(RP2D = z \mid d, m, \lambda_m) U_S(RP2D = z) \quad (4)$$

If we take expectation of  $EU_S(d, m, \lambda_m)$  over  $\mathcal{M}$  (the set of all possible  $m$ 's) and  $\Lambda_m$  (the space for  $\lambda_m$ ), we can get the expected societal utility for a design  $d$ :

$$EU_S(d) = \sum_{m \in \mathcal{M}} \pi_m(m) \int_{\Lambda_m} EU_S(d, m, \lambda_m) \pi_\lambda(\lambda_m) d\lambda_m \quad (5)$$

We can follow the steps below to approximate  $EU_S(d)$  through the Monte Carlo simulation:

For a design  $d$  and a model family pair  $m$ , do steps 1 to 6:

1. Sample  $n_\lambda$  sets of parameters from the prior distribution  $\pi_\lambda(\lambda_m)$ :

$$\{\tilde{\lambda}_m^w, w=1, \dots, n_\lambda, \tilde{\lambda}_m^w \in \Lambda_m\}$$

2. Simulate  $n_r$  clinical trials under  $d, m, \tilde{\lambda}_m^w$

3. Calculate  $\widehat{\Pr}(RP2D = z) = \frac{n(RP2D = z)}{n_r}$ ,  $z \in \mathcal{Z} \cup \mathcal{Z}^c$

4. Repeat steps 2-3 for the rest sets of sampled parameter values

5. Calculate:

$$\bar{U}_S(d, m) = \frac{1}{n_\lambda} \sum_{w=1}^{n_\lambda} \left\{ \sum_{z \in \mathcal{Z}^c} \widehat{\Pr}(RP2D = z) U_S(RP2D = z) + \sum_{z \in \mathcal{Z}} \widehat{\Pr}(RP2D = z) U_S(P_E^\dagger, P_T^\dagger \mid RP2D = z, m, \tilde{\lambda}_m^w) \right\}$$

Repeat steps 1-5 for the other possible model family pairs, then calculate  $\bar{U}_s = \sum_{m \in \mathcal{M}} \pi(m) \bar{U}_s(d, m)$ , which is the approximate estimate of  $EU_s(d)$ .

### 3.2.3 Expected total utility

We have seen that there are at least two radically distinct kinds of utility to contemplate when choosing a CT design. A third kind is the utility of total cost on a particular trial and we denote it as  $U_c$  ( $U_c < 0$ ). The total cost includes, among others, the payment to enrolled patients, the cost for dose administration, medical management, data collection, and whatever penalties accrue from delayed reporting and publishing. We assume  $U_c$  depends solely on sample size.

To make a choice of designs, one cannot safely neglect any of these three kinds of utilities, therefore, there needs to be a way to view and think about all in one context. We propose two ways to do this. The first way is to calculate the expected total personal utility ( $EU_p$ ), societal utility ( $EU_s$ ) and utility of total cost ( $EU_c$ ) for all candidate designs and summarize them in a two-way table, for example, Table 3.2. Let  $U_{Tot}$  denote total utility that encompasses total personal utility, societal utility and utility of total cost on a particular trial.  $U_{Tot}$  is a function of these three kinds of utilities whose special case is a linear function defined as:

$$U_{Tot} = \nu_p U_p + \nu_s U_s + \nu_c U_c \quad (6)$$

, where the  $\nu$ 's are parameters which convert among the different kinds of utilities. The second way is to calculate the expected total utility ( $EU_{Tot}$ ) for all candidate designs. Both ways have challenges in trading off among different kinds of utilities. The first way provides a transparent

view of the expected total personal utility, societal utility and utility of total cost for each candidate design, but CT designers have to do further analysis when no designs have the largest values across different kinds of expected utilities; the second way gives a convenient summary metric for comparing different designs, however, it may lead to a wrong decision when the justification for  $U_{Tot}$  function is questionable. In practice, CT designers may do both ways and make a choice from a careful consideration of risks and benefits.

Table 3.2 A Two-way Table for Different Kinds of Expected Utilities

		Expected Utility		
		$EU_P$	$EU_S$	$EU_C$
Design	1	1	2.5	-2000
	2	-2	4.3	-3000
	3	2	0	-1000

## **4.0 CTDESIGNEXPLORER – AN ACTION QUEUE-BASED OPEN-SOURCE SIMULATION EXPERIMENT PLATFORM FOR EVALUATING CLINICAL TRIAL DESIGNS**

### **4.1 INTRODUCTION**

In the past few decades, clinical trial (CT) designers have proposed many designs for trials at different stages of drug development. This abundance of designs mandates a question for the investigators and statisticians planning a trial: what design would be the “best” for their trial? Determining the answer must begin with careful consideration of the sometimes conflicting criteria (e.g. number of adverse events vs. number of efficacy responses) by which to judge designs.

Evaluating CT designs via simulation is used by academic research centers and pharmaceutical companies to improve the efficiency and accuracy of drug development [1-4]. Sophisticated commercial software for CT simulation and evaluation is available for those with resources to cover fees and with design challenges that happen to match the software’s capabilities. The source code of commercial software is proprietary, so users cannot easily verify the software does what it claims. Academic research centers usually use locally developed software mainly due to cost and flexibility considerations. The cost issue is illustrated by the recent quote price of a one-year single academic license for the Pharsight trial simulator:

\$11,235. Flexibility makes it possible to explore novel designs, models and evaluation criteria. Software developed locally focuses on answering specific research questions in compressed time, and therefore has limited capabilities. The open-source software development approach has helped produce reliable, high quality software quickly and inexpensively. To our knowledge, MSToolkit is the only currently available open-source software for general-purpose CT design evaluation via simulation. It provides flexible data handling, however, it also falls short in capabilities; for example, it cannot evaluate designs with adaptive treatment allocations in its most recent version 2.0.

The lack of a common framework for describing CT environments and standards for coding leads to information and software resource silos, constricting exchanges of ideas and data between innovative CT designers and those selecting existing designs for CTs. Lack of interoperability is a root cause of inefficient design evaluation processes and inconsistent evaluation results.

Therefore, we seek to build a transparent, extendible simulation experiment platform and a set of standards for further development so that CT designers can evaluate available designs and/or share their innovations. Here we introduce CTDesignExplorer - an open-source platform based on an action queue, with reliance on common names of the data elements to support interoperability. Techniques in S4 classes and methods [78] are utilized to make CTDesignExplorer extendible and reusable. To facilitate the use, we have developed CTDesignExplorer as an R package. In Section 4.2, we will describe CTDesignExplorer in detail and illustrate extensibility, reuse and sharing using some examples, and in Section 4.3 we conclude with future development.

## 4.2 CTDESIGNEXPLORER

### 4.2.1 Overview

Specification of a simulation experiment requires users to specify objects in the following five categories:

#### ***Design***

A *design* is an algorithm for making decisions in a CT. The types of decisions include allocating a treatment plan to each patient cohort, switching from the first stage of a CT to the second stage (e.g. adaptive Simon two-stage design [79]), conducting assays and other data collection actions, and stopping or continuing a CT. A *design specification* includes both the type of design and a specific value for a vector of design parameters.

#### ***Baseline Characteristic Model***

A *baseline characteristic model* is a description of the sampling distribution for one baseline characteristic of a patient. A patient's baseline characteristics may be dependent on each other. To model the total joint distribution of a patient's baseline characteristics, we model each baseline characteristic using sequential conditional distributions. A *baseline characteristic model specification* includes the baseline characteristic name, the names of other baseline characteristics and a function for generating the baseline characteristic. It does not include whether and when the baseline characteristic is to be measured; instead, that is determined by the design.

#### ***Population Model***

A *population model* is a description of the joint distribution of all baseline characteristics that may affect either the patient's outcomes, times to outcomes, or patient-level decision-

making in a CT. A *population model specification* is a list of *baseline characteristic model specifications*.

### **Outcome Model**

An *outcome model* is a rule for generating a patient's outcomes and/or times to outcomes, as a function of his/her baseline characteristics, treatment plan allocations and previous outcomes. An *outcome model specification* includes the type of outcome model and specific values for a vector of model parameters.

### **Evaluation Criterion**

An *evaluation criterion* refers to an operating characteristic of a design. An *evaluation criterion specification* may include specific values for a vector of criterion parameters if any.

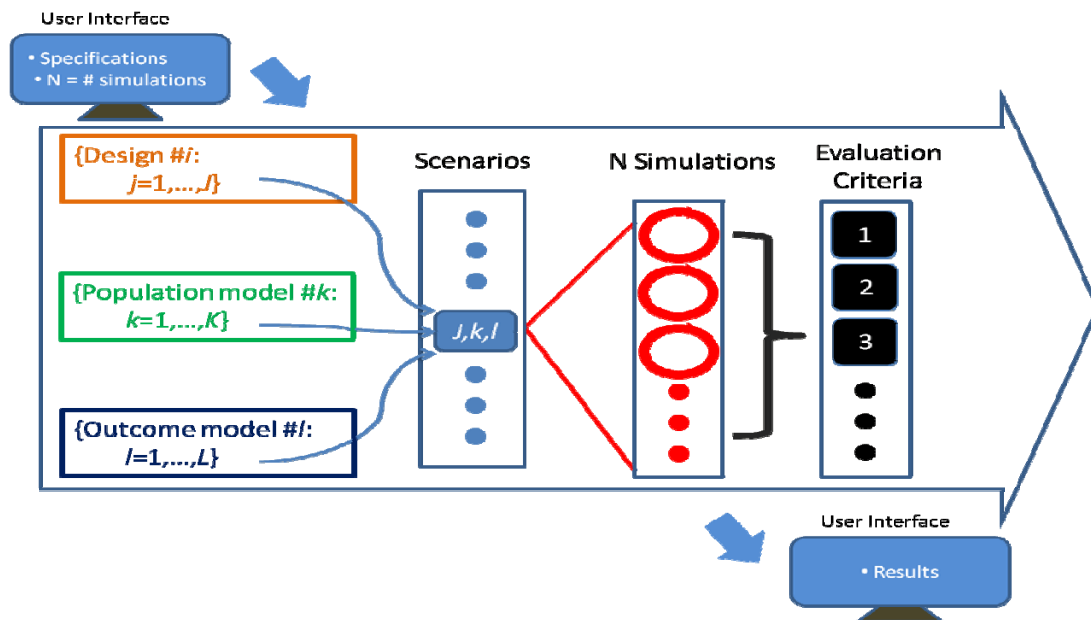


Figure 4.1 Overview of CTDesignExplorer



Figure 4.1 displays the overview of CTDesignExplorer. To perform a simulation experiment, users need to provide the specifications for objects: designs, population models, outcome models and evaluation criteria of interest. After CTDesignExplorer receives the inputs from users, including the number of CT simulations, it automatically checks what objects can interoperate with each other (interoperability among objects will be discussed in Section 4.2.3). Each combination of a design, population model and outcome model constitutes a scenario. For each scenario, CTDesignExplorer runs  $N$  CT simulations, and the simulated CT data are evaluated by the corresponding criteria. After simulation and evaluation are done for all scenarios, the evaluation results are displayed in the user interface.

Users can also save simulated data to a specified directory, together with the seed for random number generations to allow replication of results.

#### **4.2.2 Simulation framework**

Figure 4.2 displays the framework we use for simulating a single CT at a certain scenario. An action queue is a list of actions ordered by their execution times. The solid dark arrows trace simulation steps in the framework; the blue dotted arrows show action exchanges between the action queue and steps; the green dotted arrows show data exchanges between the temporary CT data repository and steps. At the start of the simulation, the framework initializes a temporary CT data repository and steps. At the start of the simulation, the framework initializes a temporary CT data repository and an action queue. The initial actions in an action queue are generated by a design object. After initialization, the “while” loop begins. On each round, the first action in the queue is retrieved, executed and removed from the queue. The consequences from executing an action may include adding new actions to the queue and/or adding new data to the CT data

repository. If the queue is now empty, the simulation stops, otherwise the “while” condition is satisfied, and the loop continues.

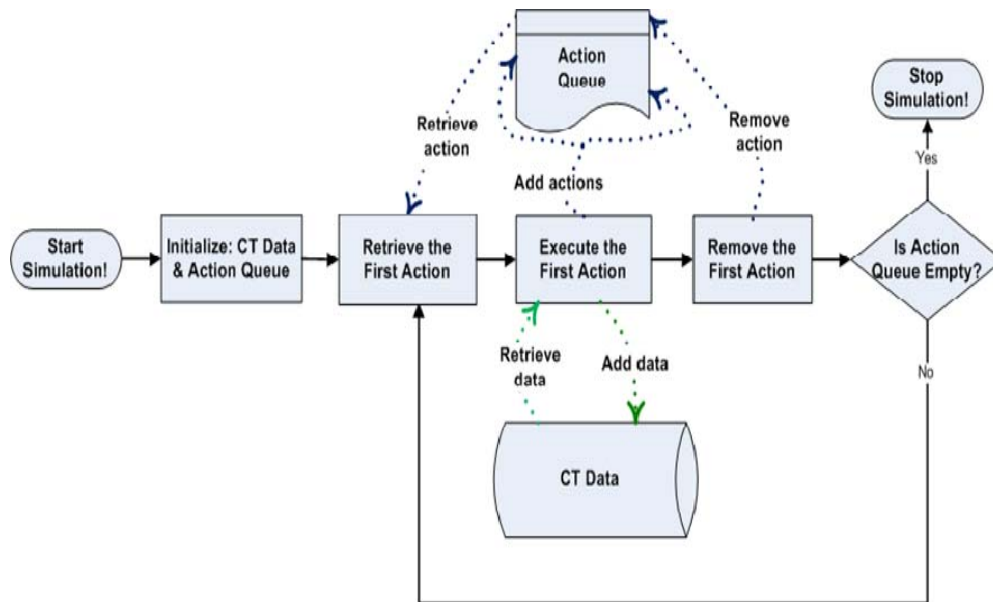


Figure 4.2 Action Queue-based Simulation Framework

The action queue-based framework is particularly useful in simulating CTs where decisions, either at the patient level or at the CT level, depend in complex ways on the current state of information. For example, in some CT designs, the protocol’s treatment of a patient may change while on study due to a crossover design element, a rule for dosage modification or delay due to toxicity, or a secondary randomization. Also, patient histories may overlap, which complicates the application of stopping rules or treatment allocation rules. The action queue handles all these cases. It adds new actions adaptively, as needed, guided by accumulating data.

### 4.2.3 Interoperability

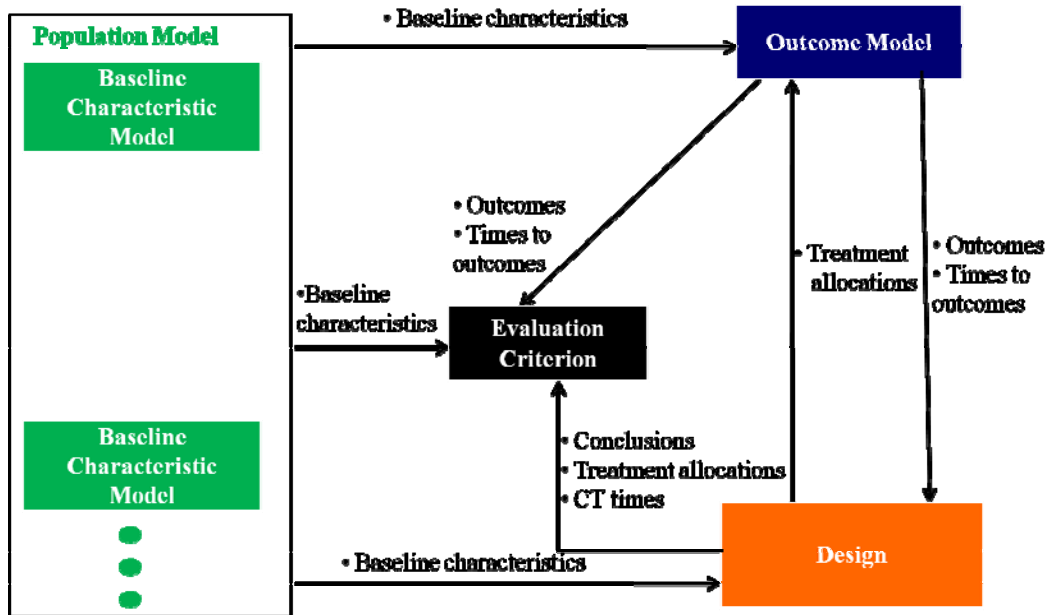


Figure 4.3 Communications among Objects

In CTDesignExplorer, objects (designs, baseline characteristic models, population models and outcome models) may require information provided by other objects to perform their functions. Thus, in our terminology, the information that an object provides is called provisions and the information that an object requires is called requirements. As displayed in Figure 4.3, baseline characteristic models are nested within a population model as a whole to communicate with other types of objects. The population model provides values of baseline characteristics to the outcome model, design and evaluation criterion; the outcome model provides outcomes (possibly including time-to-event data) to both the design and the evaluation criterion; the design provides treatment allocations to both the outcome model and the evaluation criterion, additionally, the design provides summary results (conclusions) from running a CT and CT times to the

evaluation criterion. CT time is the time at which a CT-level event occurs, for example, time when a CT ends. In Figure 4.3, arrows point to requirers and away from providers.

We have providers and requirers communicate using common names of the data elements. Examples of these data elements along with their common names are presented in Table 4.1.

We use a name matching mechanism to check interoperability between requirers and providers. For example, if a design requires “BinaryToxicity” outcomes and an outcome model provides “ToxicityGrade” outcomes, this design does not interoperate with this outcome model, because it requires binary toxicity outcomes and the outcome model provides ordinal toxicity outcomes.

Table 4.1 Examples of Common Names for Data Elements

<b>Provided by</b>	<b>Type</b>	<b>Name</b>	<b>Definition</b>	<b>PossibleValues</b>
Population Model	Baseline characteristic	SubPopIndex	Which sub-population a patient belongs to	1, 2, 3...
Design	An element of a treatment allocation	Dose	The dosage of a treatment	$\geq 0$
Design	Conclusion	RP2D	Recommended Phase II dose	$> 0$ , NA
Design	CT time	CTEndTime	Time when a CT ends	$\geq 0$ , NULL
Outcome Model	Outcome	BinaryToxicity	Indicator whether a patient experiences a dose-limiting toxicity (DLT)	0, 1
Outcome Model	Outcome	ToxicityGrade	Ordinal toxicity outcome, as a grade	1,2,3,4,5
Outcome Model	Time to outcome	TimeToToxicity	Time to a DLT event	$\geq 0$

To facilitate use of common names of the data elements and to promote contributions from a wide community of CT designers, we will develop similar interactive tools to “CDE

Browser” and “CDE Curation Tool” used in the cancer Data Standards Registry and Repository (caDSR) for browsing and managing the common data elements.

#### **4.2.4 Implementation: S4 classes and methods**

We designed and implemented CTDesignExplorer using the object oriented programming (OOP) paradigm. OOP encapsulates the representation of objects, which helps to modularize a complex software system; it also provides class inheritance and method-dynamic dispatch, which makes it possible to both build a common framework and extend the software for innovations. The code base relies on S4 classes and methods [78] within R. S4 OOP is slightly different from traditional OOP other languages like Java and Python follows, in that the method definitions in the S4 system do not reside in a class definition. Instead, methods having the same name are stored within the same generic function according to their signature, a named list of classes with the names corresponding to the arguments’ names of the generic function.

An S4 class is declared by a call to `setClass`, along with the named slots which contains relevant information. The instances of a class are validated against its definition. An S4 method cannot be declared by a call to `setMethod` unless its corresponding generic function has been declared using the `setGeneric` function. In the following, we will show the representations of the objects in CTDesignExplorer by S4 classes and methods.

For baseline characteristic models, the class “BaseCharModelSpecifier” represents their specifications and an associated method for generating a single patient’s baseline characteristic.

For population models, the class “PopModelSpecifier” represents their specifications, and there are two associated methods. One method sequentially generates each of a patient’s baseline

characteristic values, and the other method generates these values for a specified number of patients.

For a specific type of designs, a subclass of the virtual class “DesignSpecifier” represents their specifications, and there are associated methods for each type of decision, such as allocating a set of concurrent treatments.

For a specific type of outcome models, a subclass of the virtual class “OutcomeModelSpecifier” represents their specifications, and there are two associated methods. One method generates a patient’s outcomes and/or times to outcomes from a set of concurrent treatments, and the other method generates outcomes and/or times to outcomes from a set of concurrent treatments for a specified group of patients.

For a specific type of evaluation criteria, a subclass of the virtual class “EvalSpecifier” represents their specifications, and there is one associated method for evaluating a design under this type of criteria.

Additionally, there are methods for getting provisions and/or requirements from designs, baseline characteristic models, population models, outcome models and evaluation criteria.

Data from a single CT simulation are represented by the class “CTData”. Figure 4.4 illustrates the hierarchical data structure. Each patient’s data is represented by the class “PatData”, and data from each set of concurrent treatments are represented by the class “ConcurrentTrtsData”. A default object instantiated from the class “CTData” has an empty list for “PatsData” slot, values of NULL for both CT times and conclusions.

Actions in an action queue are represented by the class “Action”. Each action object contains information for calling the corresponding method and the time when it is executed. An example of the corresponding method to an action is `allocateTrts`, which returns a named

list of two elements: one is the updated CT data with new treatment allocations information for a specified group of patients, and the other is a list of new actions (e.g. a new action can be generating this group of patients' outcomes after the treatment). Only methods associated with designs can be action methods. Action queues are represented by the class "ActionQueue". Its associated method `addAction` inserts a new action to the right position of the queue, and returns the updated action queue.

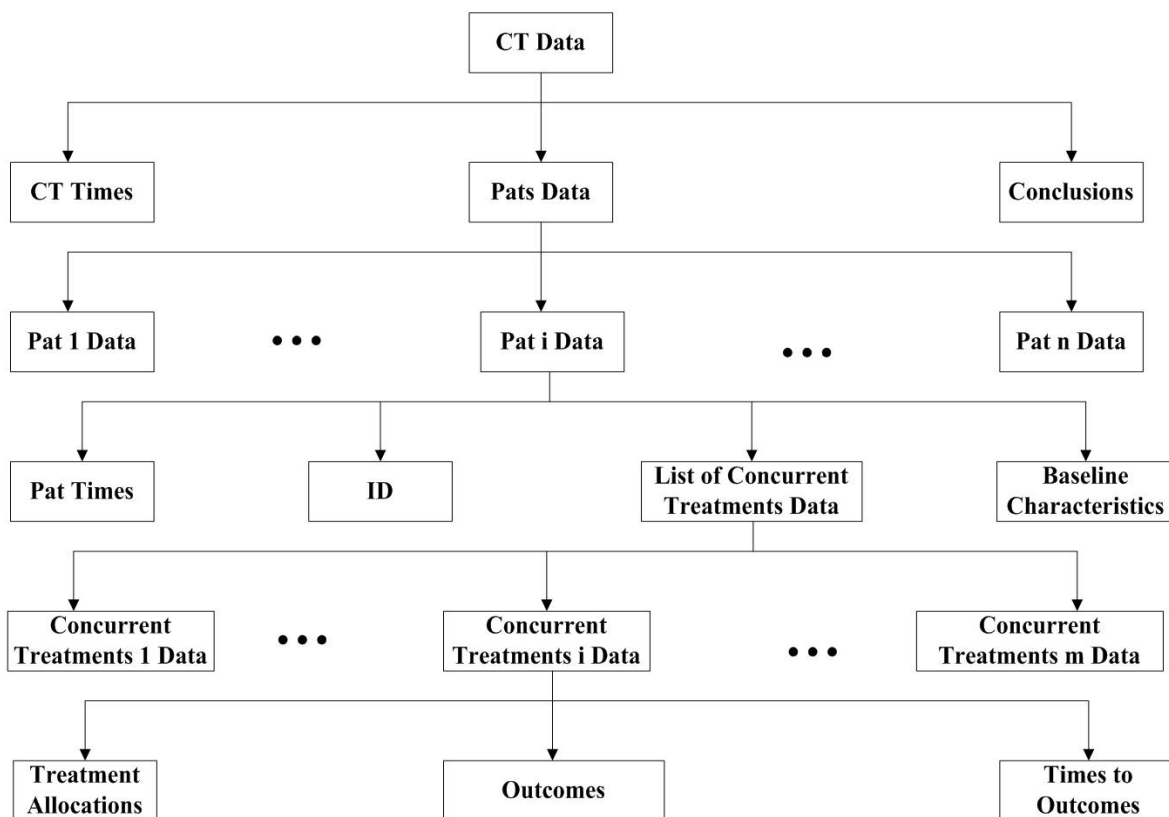


Figure 4.4 CT Data Structure

#### 4.2.5 Regular use: using existing classes and methods

Regular use refers to using only already existing classes and methods in the CTDesignExplorer release. The essence of the flexibility in evaluating designs using our platform is that it allows for different population models, outcome models and evaluation criteria. We list below some use cases of interest to colleagues. The uses of CTDesignExplorer are not limited to these examples.

- A protocol statistician worries about the potential adverse effect of patient heterogeneity in the accrual pool on the performance of a Phase I dose-finding design. The statistician specifies population models for a Phase I patient population with and without heterogeneity respectively; specifies the Phase I trial design he/she cares about; specifies a plausible underlying true outcome model; specifies the evaluation criteria that are used to assess performance of the design; runs the experiment with these inputs and then compares the evaluation results between the scenarios using different population models. This use case is demonstrated in Section 5.2.
- An investigator has concerns about the cost-effectiveness of genotyping patients for identifying a certain allele which may affect a patient's outcome on treatment due to pharmacokinetic or pharmacodynamic effects. The investigator specifies a population model for a patient population with a specified proportion of patients having that allele; specifies the two designs, one accounts for and the other disregards that genetic information; specifies a plausible underlying true outcome model in which whether or not having that gene is a dichotomized covariate; specifies the evaluation criteria that can be used to estimate the cost-effectiveness, such as expected total utility; runs the experiment with these inputs and compare the evaluation results between two designs.



- A research statistician is interested from a methodological perspective in investigating how different choices of prior distribution for parameters in the assumed logit dose-toxicity model affect the performance of CRM designs. He/she specifies a population model for a patient population if necessary; specifies the CRM designs which are only different with respect to the prior distributions; specifies several true outcome models; specifies the evaluation criteria to assess the performance of the designs; runs the experiment with these inputs and compares the evaluation results between these designs. This use case is demonstrated in Section 5.1.4.

Users do not need to know S4 classes and methods in order to use CTDesignExplorer and conduct simulation experiments. We provide the interfacing function “specifyObject” which allows users to specify a baseline characteristic model, population model, outcome model, design and evaluation criterion using a regular R function. The following shows an example on how to specify a standard Phase I “3+3” design:

First, find all available classes for representing a design specification

```
> subClassNames("DesignSpecifier")
[1] "APlusBSpecifier"
[2] "CRMSpecifier"
[3] "Phase2BryantDaySpecifier"
```

The “APlusBSpecifier” class represents specifications for “A+B” with dose de-escalation designs [75], of which the “3+3” design is a special case.

Second, obtain documentation on class “APlusBSpecifier”

```
> class?APlusBSpecifier
```

An HTML window will pop up for the documentation part of which is shown in Figure 4.5.

Finally, specify a “3+3” design

```
> specifyObject(
+   className="APlusBSpecifier",
```

```

+     slots=list(A=3,B=3,C=1,D=1,E=1,TierDoses=c(3.0,6.0,9.9,15.0
+           ,21.1,28.0)
+ )
An object of class "APlusBSpecifier"
Slot "A":
[1] 3
Slot "B":
[1] 3
Slot "C":
[1] 1
Slot "D":
[1] 1
Slot "E":
[1] 1
Slot "TierDoses":
[1] 3.0 6.0 9.9 15.0 21.1 28.0

```

## Class "APlusBSpecifier"

## Description

An S4 class, representing the specifications for "A+B" with dose de-escalation designs.

## Objects from the Class

Objects can be created by calls of the form `new("APlusBSpecifier", ...)`.

## Slots

A:

Object of class "numeric", the initial cohort size.

B:

Object of class "numeric", the additional cohort size.

C, D, E:

Objects of class "numeric", binary toxicity (e.g. dose-limiting toxicity, DLT) counts associated with stopping the trial and dose assignment for the next group of patients.

TierDoses:

Object of class "numeric", a vector of selected doses for the testing.

## Extends

Class "[DesignSpecifier](#)", directly.

## Methods

Figure 4.5 Partial Documentation for Class "APlusBSpecifier"

## 4.2.6 Advanced use: extending, reusing and sharing

Advanced use refers to extending our platform, reusing code for already existing specifierObject classes, and sharing innovations with a community of CT designers.

### 4.2.6.1 Extending

Since we allow users to write their own functions for generating a baseline characteristic when specifying a baseline characteristic model, CTDesignExplorer can cover any user-defined baseline characteristic model and population model. Users can extend CTDesignExplorer for their novel designs, outcome models and evaluation criteria by developing new subclasses and the associated methods. We provide templates for users to follow when extending the software for different types of innovations. For example, the following lists the templates for extending CTDesignExplorer for a new design. The bold italics and “xxx” in the templates will be substituted by users.

- To develop a new subclass of class “DesignSpecifier” for the new design specification, this template is used:

```
setClass("NewDesignSpecifier",  
  representation(DesignParam1=xxx, DesignParam2=xxx, xxx),  
  contains="DesignSpecifier",  
  prototype=list(xxx),  
  validity=function(object){xxx})
```

For example, if this new design is a CRM design [6], “**NewDesign**” in the above template can be substituted by “CRM”.

- To develop associated methods for the new design, these templates are used:

```
setMethod("generateInitialActions",  
  signature(designSpec="NewDesignSpecifier"),  
  function(designSpec){xxx})
```

The above method returns a list of initial actions.

The returned value from each of the methods below is a named list with at most two elements: the element with name “NewCTData” stores the updated CT data; and the element with name “NewActions” stores a list of new actions.

```
setMethod("allocateTrts",  
  signature(designSpec="NewDesignSpecifier",  
    currentCTData="CTData",  
    currentGlobalTime="numeric",  
    patsIndices="numeric"),  
  function(designSpec, currentCTData, currentGlobalTime, patsIndices) {xxx})
```

The above method returns a named list with two elements.

```
setMethod("checkStoppingRule",  
  signature(designSpec="NewDesignSpecifier",  
    currentCTData="CTData",  
    currentGlobalTime="numeric"),  
  function(designSpec, currentCTData, currentGlobalTime) {xxx})
```

The above method returns a named list with one element; that element’s name is either “NewCTData” (if the stopping rule decides to stop the CT) or “NewActions” (if the stopping rule decides to continue the CT).

```
setMethod("checkSwitchingStageRule",  
  signature(designSpec="NewDesignSpecifier",  
    currentCTData="CTData",  
    currentGlobalTime="numeric"),  
  function(designSpec, currentCTData, currentGlobalTime) {xxx})
```

The above method is optional, depending on whether the new design is multiple-stage design. It returns a named list with updated CT data. If the switching-stage rule decides to switch the current stage of a trial to a higher stage, the “SwitchingStageTime” would be changed from NULL to the current global time.

The new design may have more associated methods, e.g. for checking off-CT rule. Then users need to develop a generic function first and then the method.

- To develop methods for getting provisions and requirements of the new design, the following two templates are used.

```
setMethod("getProvisions",
  signature(spec="NewDesignSpecifier"),
  function(spec) {xxx})
```

The above method returns a named list with elements for the names of treatment allocation elements, CT times and conclusions that the new design provides.

```
setMethod("getRequirements",
  signature(spec="NewDesignSpecifier"),
  function(spec) {xxx})
```

The above method returns a named list with elements for the names of baseline characteristics, outcomes and times to outcomes that the new design requires.

#### 4.2.6.2 Reusing

The class inheritance central to object-oriented programming greatly facilitates the reuse of code for already existing specifierObject classes in the CTDesignExplorer release. Subclasses of a class inherit all the methods associated with that class, and R provides mechanisms to select a method based on inheritance distance[78]. For example, if the algorithm of the new design for allocating a set of concurrent treatments is the same as that of an existing design, instead of developing a subclass of “DesignSpecifier” in the first template above, users can reuse the code for implementing that algorithm by developing a subclass of the class for the existing design specification.

#### 4.2.6.3 Sharing

Users have the following two options for sharing their innovations after developing new classes and associated methods:

- Submit their codes to us, and we will include them in the next version of CTDesignExplorer if the submission meets a set of requirements (e.g. requirements for validation and documentation)
- Develop a small new R package which depends on CTDesignExplorer, to be released independently.

#### 4.2.7 Code Validation

Code validation after implementation is an important step to assure users of the quality of software. In the spirit of open-source software (OSS) – “given enough eyeballs, all bugs are shallow” [80], the intent and hope is to rely heavily on peer code review for an overall improvement of the quality of software. To facilitate this process, we plan to register a project for CTDesignExplorer on R-Forge [81] which offers a central platform for the development of R packages, R-related software and other projects. R-Forge provides quality management system similar to that of CRAN which checks packages based on `R CMD check` at least once daily; it also provides bug tracking system which allows users to notify package authors about the problems they encounter.

Before we release new version of CTDesignExplorer, we will validate code for new subclasses and methods provided by either us or contributors. Two ways can be used for validating the code for new subclasses and methods associated with outcome models and evaluation criteria: one way is to implement new outcome models in S3 system and to compare results obtained from the two kinds of implementation; another way is to ask a third person to examine the code.

For new subclasses and methods associated with a design, we will first find other software which implements the same design. If the other software is available, we will compare two things: one replication of the simulated clinical trial data and evaluation results under certain criteria, obtained from the other software with those obtained from the new subclasses and methods. If the other software is not available, we will ask a third person either to examine the code or to write code for implementing such design in S3 system and then we will compare his/her simulated data and evaluation results with those from the new subclasses and methods. In the following, we will give two examples demonstrating code validation for subclasses and methods associated with designs.

#### **4.2.7.1 Code validation for “A+B” design**

Yong Lin has written the “pmtd” program for numeric calculation of the evaluation criteria discussed in his paper “Statistical Properties of the Traditional Algorithm-based Designs for Phase I Cancer Clinical Trials”[75]. The ‘pmtd’ program is available from “<http://www2.umdj.edu/~lino/>”. Even though the “pmtd” is an S-plus program, we can run it without any modifications in the R environment. The “A+B” design implemented in CTDesignExplorer corresponds to the “A+B” design with dose de-escalation in Lin’s paper. We did not find publicly available software for simulating data from clinical trials using the “A+B” design.

To compare our evaluation results with the analytical results from the “pmtd” program, we performed a simulation experiment for the standard “3+3” design. The experiment is set up as below:

*Design:* the standard “3+3” design



*Population Model:* None. We assume there are no individual's baseline characteristics that will affect the patients' outcomes.

*Outcome Model:* a non-parametric dose-toxicity model. The corresponding toxicity probabilities to the testing tier doses are listed in the table below:

Table 4.2 The Outcome Model Used in the Code Validation for "A+B" Design

Tier Dose	3.0	6.0	9.9	15.0	21.1
Pr(T)	0.05	0.1	0.15	0.25	0.4

*Evaluation Criteria:*

- Probability of each tier dose being chosen as RP2D
- Average number of patients allocated at each tier dose
- Average number of toxicities observed at each tier dose

*Number of Simulations:* 1000

In the next, we will look at the comparison results. For simplicity, we abbreviate CTDesignExplorer as CTDE in the following tables within this section and Section 4.2.7.2.

We denote the chosen RP2D as  $NA_i$  if dose escalation is still indicated at the highest tier dose, and as  $NA_r$  if dose de-escalation is indicated at the lowest tier dose. Table 4.3 shows the first comparison result on the probability of each tier dose being chosen as RP2D; Table 4.4 shows the second comparison result on the average number of patients allocated at each tier dose; and Table 4.5 shows the third comparison result on the average number of toxicities observed at each tier dose.

Table 4.3 First Comparison Result in the Code Validation for “A+B” Design

	<b>RP2D</b>						
	<b>3.0</b>	<b>6.0</b>	<b>9.9</b>	<b>15.0</b>	<b>21.1</b>	$NA_t$	$NA_r$
<b>CTDE</b>	0.081	0.178	0.321	0.283	0	0.116	0.021
<b>Pmtd</b>	0.095	0.175	0.305	0.265	0	0.133	0.027
<b>(CTDE –pmtd)<sup>2</sup>/pmtd<sup>2</sup></b>	0.022	<0.001	0.003	0.005	NA	0.016	0.049
<b>Chi-Squared Test</b>	<ul style="list-style-type: none"> <li>The test is comparing the observed frequencies of each tier dose being chosen as RP2D except for the dose of 21.1 in CTDE against the calculated probabilities from pmtd.</li> <li>P value: 0.17</li> </ul>						

Table 4.4 Second Comparison Result in the Code Validation for “A+B” Design

	<b>Tier Dose</b>				
	<b>3.0</b>	<b>6.0</b>	<b>9.9</b>	<b>15.0</b>	<b>21.1</b>
<b>CTDE</b>	3.630	4.107	4.347	3.825	1.803
<b>pmtd</b>	3.658	4.062	4.231	3.689	1.850
<b>(CTDE –pmtd)<sup>2</sup>/pmtd<sup>2</sup></b>	<0.001	<0.001	0.001	0.001	0.001
<b>Chi-Squared Test</b>	<ul style="list-style-type: none"> <li>The test is comparing the average number of patients allocated at each tier dose in CTDE against the calculated probabilities from pmtd.</li> <li>P value: 0.9994</li> </ul>				

Table 4.5 Third Comparison Result in the Code Validation for “A+B” Design

	<b>Tier Dose</b>				
	<b>3.0</b>	<b>6.0</b>	<b>9.9</b>	<b>15.0</b>	<b>21.1</b>
<b>CTDE</b>	0.132	0.384	0.658	0.956	0.754
<b>pmtd</b>	0.183	0.406	0.635	0.922	0.740
<b>(CTDE –pmtd)<sup>2</sup>/pmtd<sup>2</sup></b>	0.078	0.003	0.001	0.001	<0.001
<b>Chi-Squared Test</b>	<ul style="list-style-type: none"> <li>The test is comparing the average number of toxicities observed at each tier dose in CTDE against the calculated probabilities from pmtd.</li> <li>P value: 0.93</li> </ul>				

We found from the above comparisons that the evaluation results for the standard “3+3” design using CTDesignExplorer are very similar to those analytically calculated results from “pmtcd” program.

#### 4.2.7.2 Code validation for the CRM

We compared one replication of CT data (patient Level: binary toxicity outcome and the assigned dose level; CT level: recommended Phase 2 dose level (RP2DL)) and evaluation results from 1000 simulations with those obtained from another R package “dferm”, which can be downloaded from CRAN.

Specifically, we consider 16 different CRM designs as described in the table below:

Table 4.6 Designs Used in the in the Code Validation for the CRM

Design #	Model Type	Restriction	# Stage	Cohort Size
1	Exponential	Yes	One	1
2	Exponential	Yes	One	2
3	Exponential	Yes	Two	1
4	Exponential	Yes	Two	2
5	Exponential	No	One	1
6	Exponential	No	One	2
7	Exponential	No	Two	1
8	Exponential	No	Two	2
9	Logit	Yes	One	1
10	Logit	Yes	One	2
11	Logit	Yes	Two	1
12	Logit	Yes	Two	2
13	Logit	No	One	1
14	Logit	No	One	2
15	Logit	No	Two	1
16	Logit	No	Two	2

In the above table, *Model Type* refers to type of the assumed one-parameter dose-toxicity model; *Restriction* refers to the dose escalation restriction proposed by Cheung in his “dfcrm” package which avoids (1) skipping doses in escalation and (2) escalating when the proportion of patients experiencing toxicity in the last patient cohort is larger than or equal to the target toxicity rate (i.e., incoherent escalation); two-stage CRM includes an initial stage where the dose level is escalated according to a pre-specified dose level sequence.

These 16 CRM designs share the following design parameters:

- Initial toxicity probability guesses: 0.04, 0.12, 0.16, 0.23, 0.44
- Tier doses: 3.0, 6.0, 9.9, 15.0, 21.1
- Target toxicity rate: 0.15
- Sample Size: 24
- Starting dose level if the design is one-stage CRM: 2
- Dose level assignment sequence in the initial stage if the design is two-stage CRM are: 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5.
- The intercept value for the logit model = 3
- The prior distribution is normal with mean 0 and standard deviation 1.34

The population and outcome models, and evaluation criteria are the same as those in Section 4.2.7.1.

The R functions used for generating random numbers in the CRM design implementation are the same between CTDE and “dfcrm”, and we use the same initial seed for random number generations when running experiments. Assuming “dfcrm” code is valid, we expect the one replication of CT data and evaluation results from 1000 simulations obtained from CTDE would be the same as those from “dfcrm” if CTDE code is valid.

We will now look at the simulation results obtained from CTDE and “dferm” when CT design is design 1:

- One replication of CT data

Table 4.7 First Comparison Result in the Code Validation for the CRM

Patient ID	CTDE		dferm	
	Toxicity	Assigned Dose Level	Toxicity	Assigned Dose Level
1	0	2	0	2
2	1	3	1	3
3	0	1	0	1
4	0	1	0	1
5	1	1	1	1
6	0	1	0	1
7	0	1	0	1
8	0	1	0	1
9	0	1	0	1
10	0	1	0	1
11	1	1	1	1
12	0	1	0	1
13	0	1	0	1
14	0	1	0	1
15	0	1	0	1
16	0	1	0	1
17	0	1	0	1
18	0	1	0	1
19	0	1	0	1
20	0	1	0	1
21	0	1	0	1
22	0	1	0	1
23	0	1	0	1
24	0	1	0	1
<b>RP2DL</b>	1		1	

- Probability of each tier dose being chosen as RP2D

Table 4.8 Second Comparison Result in the Code Validation for the CRM

	<b>RP2D</b>				
	<b>3.0</b>	<b>6.0</b>	<b>9.9</b>	<b>15.0</b>	<b>21.1</b>
<b>CTDE</b>	0.109	0.307	0.308	0.248	0.028
<b>dfcrm</b>	0.109	0.307	0.308	0.248	0.028

- Average number of patients allocated at each tier dose

Table 4.9 Third Comparison Result in the Code Validation for the CRM

	<b>Tier Dose</b>				
	<b>3.0</b>	<b>6.0</b>	<b>9.9</b>	<b>15.0</b>	<b>21.1</b>
<b>CTDE</b>	4.382	6.053	5.582	5.72	2.263
<b>dfcrm</b>	4.382	6.053	5.582	5.72	2.263

- Average number of toxicities observed at each tier dose

Table 4.10 Fourth Comparison Result in the Code Validation for the CRM

	<b>Tier Dose</b>				
	<b>3.0</b>	<b>6.0</b>	<b>9.9</b>	<b>15.0</b>	<b>21.1</b>
<b>CTDE</b>	0.21	0.573	0.849	1.439	0.898
<b>dfcrm</b>	0.21	0.573	0.849	1.439	0.898

The above tables show no difference in both the simulated data and evaluation results between CTDE and “dfcrm”. We did not observe differences for the other 15 designs as well (we omitted the tables here).

### **4.3 FUTURE DEVELOPMENT**

We present an open-source simulation experiment platform for evaluating competing CT designs. An R package called CTDesignExplorer has been developed based on the source code as of April 30th, 2010. Future development will include setting up requirements for contributors, publishing CTDesignExplorer, project registration on R-Forge to facilitate collaborative software development and building a user-friendly GUI. Complex factorial evaluations will be computationally time-consuming. Parallelization for these demanding tasks is possible, for example, several R packages provide support for parallel processing [82].

## **5.0 SOFTWARE APPLICATION IN EVALUATING EARLY PHASE CLINICAL TRIAL DESIGNS**

Simulation allows CT designers to assess the consequences of the design factors and the assumptions made. The purpose of this chapter is to present several examples of using the platform to investigate important issues in model formulation, choice of prior distributions, and patient heterogeneity for early phase clinical trial designs. Specifically, we will look at the logit model in the continual reassessment method (CRM), choices of prior distribution for its model parameters, and the effect of patient heterogeneity on the performance of the standard “3+3” design and the CRM.

### **5.1 LOGIT MODEL IN THE CONTINUAL REASSESSMENT METHOD**

#### **5.1.1 The continual reassessment method (CRM)**

The CRM was first introduced by O’Quigley et al. in the year 1990 [6], and has drawn much attention from the biostatistical community [7-14]. The CRM assumes that the probability of toxicity response increases monotonically with increasing dose via a parametric model  $p_j = p(z_j^* | \boldsymbol{\phi})$ , where  $j$  is the index for tier dose level,  $\boldsymbol{\phi}$  is a vector of parameters that can be of



length one or more than one,  $p_j$  is the toxicity probability at tier dose level  $j$ , and  $z_j^*$  is very rarely the actual therapeutic dose, but rather is a rescaled dose at level  $j$ , calculated from the initial guesses of toxicity probabilities ( $\pi_j, j=1, \dots, J$ ) and prior means of parameters ( $\phi_0$ ). Substituting  $p_j$  by  $\pi_j$ , and  $\phi$  by  $\phi_0$  into the chosen parametric model, we can find a solution for  $z_j^*$  that is assigned to tier dose level  $j$ . Commonly used models include the conventional logit model (“conventional” is used to distinguish this model from the re-parameterized logit model to be introduced in Section 5.1.2)

$$\log\left(\frac{p_j}{1-p_j}\right) = \alpha + \exp(\beta)z_j^* \quad (7)$$

, and the exponential model:

$$p_j = (z_j^*)^{\exp(\beta)} \quad (8)$$

In CRM, the dose-toxicity response relationship is continually re-assessed based on accumulating data collected from the trial. The next patient cohort is typically assigned the dose which has the posterior toxicity probability closest to the target toxicity rate, although the next patient cohort can receive the highest dose with the posterior toxicity probability lower or equal to the target toxicity rate in some trials for safety concerns. Variations of the CRM include using different clinical trial stopping rules, applying dose-escalation restrictions, different number of patients per cohort, and different starting dose levels.

### 5.1.2 Re-parameterized logit model

In the original CRM and its variations, single-parameter dose-toxicity response models are usually used. For example, in the conventional logit model,  $\beta$  is the free parameter, and

$\alpha$  typically takes a fixed value of 3. Opinion has diverged about the number of parameters to use in the model [16-20]. Supporters for single-parameter models argue that dose-toxicity response relationship can be adequately approximated by a single parameter when only focusing in the range of the true target dose and that it is not possible to reliably estimate a large number of parameters in Phase I clinical trials, which provide only a small amount of information. However, as Phase I trials grow in complexity, for example, accounting for late-onset toxicity response and combined therapy, single-parameter models face a great challenge to describe the underlying outcome model adequately. Furthermore, a recent simulation study performed by Gerke and Siedentop [16] showed that the use of a two-parameter model may lead to improved design performance compared to the use of a one-parameter model.

In the dose-toxicity response models for the CRM design, the use of rescaled doses obscures the interpretations of parameters, although it may promote a reasonable fit to models [7] or ease computation[6]. For example, in the conventional logit model, how are  $\alpha$  and  $\beta$  associated with the assumed dose-toxicity response curve? Understanding the interpretations of parameters helps to determine the number of free parameters in a model, and set up sensible prior distributions that genuinely reflect prior belief.

In the following, we will describe step by step how we developed the re-parameterized logit model which has more interpretable parameters than those in the conventional model, and then show some examples of the implications for the prior distribution of the toxicity probability at the starting dose level when  $\alpha$  in the conventional logit model takes different fixed values and prior distributions.

Suppose that  $\alpha_0$  and  $\beta_0$  are the prior means for  $\alpha$  and  $\beta$  respectively (if  $\alpha$  is not a free parameter,  $\alpha_0$  would be the fixed value for  $\alpha$ ). According to the definition of rescaled doses, substitute  $p_j$  by  $\pi_j$ ,  $\alpha$  by  $\alpha_0$ , and  $\beta$  by  $\beta_0$  in Equation (7), we can get:

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \alpha_0 + \exp(\beta_0)z_j^* \quad (9)$$

From Equation (9), we can derive  $z_j^* = \exp(-\beta_0)[\log(\frac{\pi_j}{1-\pi_j}) - \alpha_0]$ , plug it into Equation (7), we get:

$$\log\left(\frac{p_j}{1-p_j}\right) = \alpha + \exp(\beta - \beta_0)[\log\left(\frac{\pi_j}{1-\pi_j}\right) - \alpha_0] \quad (10)$$

Substituting index  $j$  by  $i$  in the above equation, we have:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \exp(\beta - \beta_0)[\log\left(\frac{\pi_i}{1-\pi_i}\right) - \alpha_0] \quad (11)$$

Subtracting (11) from (10), we get:

$$\log\left(\frac{p_j}{1-p_j}\right) - \log\left(\frac{p_i}{1-p_i}\right) = \exp(\beta - \beta_0)[\log\left(\frac{\pi_j}{1-\pi_j}\right) - \log\left(\frac{\pi_i}{1-\pi_i}\right)] \quad (12)$$

$\log\left(\frac{p_j}{1-p_j}\right) - \log\left(\frac{p_i}{1-p_i}\right)$  is the log odds ratio of having toxicity response between dose levels  $j$  and  $i$ , and  $\log\left(\frac{\pi_j}{1-\pi_j}\right) - \log\left(\frac{\pi_i}{1-\pi_i}\right)$  is the initially guessed log odds ratio for these two dose levels. Since  $j$  and  $i$  are arbitrarily chosen from  $1, 2, \dots, J$ , the conventional logit model actually assumes that for any pair of dose levels, the ratio of true log odds ratio and the initially guessed log odds ratio is the same and positive.

Let  $p_s$  and  $\pi_s$  are the true and initially assumed toxicity probability at the starting dose level respectively, plug them into Equation (12), add  $\log(\frac{p_s}{1-p_s})$  to the two sides of the equation, and substitute  $\exp(\beta - \beta_0)$  by  $\gamma$ , which is the ratio of true log odds ratio and the initially guessed log odds ratio for any pair of dose levels, then we get the re-parameterized logit model:

$$\log\left(\frac{p_j}{1-p_j}\right) = \log\left(\frac{p_s}{1-p_s}\right) + \gamma \left[ \log\left(\frac{\pi_j}{1-\pi_j}\right) - \log\left(\frac{\pi_s}{1-\pi_s}\right) \right] \quad (13)$$

Arguably  $p_s$  and  $\gamma$  in the re-parameterized model are more interpretable than  $\alpha$  and  $\beta$ . Figure 5.1 shows the relationship of  $p_s$  and  $\gamma$  to the assumed true dose-toxicity curve. In this example,  $\pi_s = 0.05$ , and the starting dose level is the lowest dose level. Red open circles correspond to the initially guessed toxicity probabilities; open circles in other colors correspond to the assumed true toxicity probabilities when  $p_s$  and  $\gamma$  take different values. From Equation (13), we know that all  $p_j = \pi_j$  if  $p_s = \pi_s$  and  $\gamma = 1$ , which implies that the initially guessed dose-toxicity curve (red curve) is the same as the assumed true dose-toxicity curve when  $p_s = \pi_s$  and  $\gamma = 1$ . We observe that when we keep  $\gamma = 1$  and change  $p_s$ , the assumed true curves have the same shape as the initially guessed and the smaller  $p_s$ , the lower assumed true curve; if we keep  $p_s = 0.05$  and change  $\gamma$ , the assumed true curves will start from the same point as the initially guessed but different shapes. We also observe that the smaller  $\gamma$ , the more slowly the assumed true curve increases with the increase of dose levels. Therefore,  $p_s$  and  $\gamma$  are interpretable with respect to their association with the assumed true dose-toxicity curve, which helps to set up prior distributions that genuinely reflect investigators' prior belief.

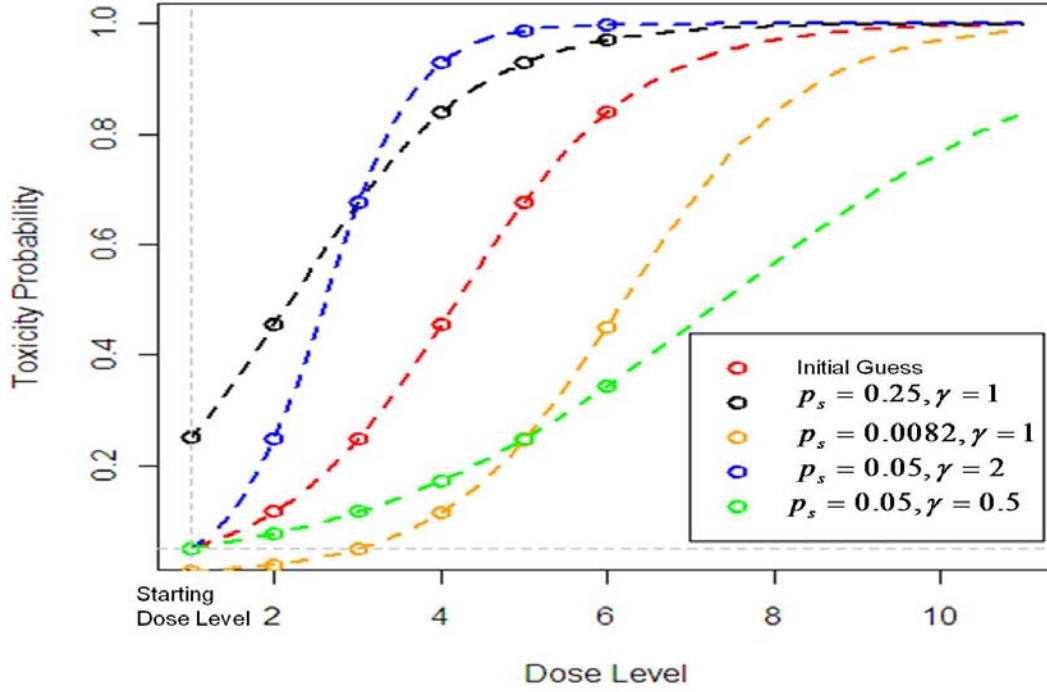


Figure 5.1 The relationship of  $p_s$  and  $\gamma$  to the Assumed Dose-toxicity Curve

The transformations between  $(p_s, \gamma)$  and  $(\alpha, \beta)$  are shown in the following two equations:

$$p_s = \frac{1}{1 + \exp(-[\alpha + \exp(\beta - \beta_0)[\log(\frac{\pi_s}{1 - \pi_s}) - \alpha_0]])} \quad (14)$$

$$\gamma = \exp(\beta - \beta_0) \quad (15)$$

Equation (14) shows that the toxicity probability at the starting dose level is a function of  $\alpha$ ,  $\beta$ , prior means and its initial guess, and Equation (15) shows that  $\gamma$  is only related with  $\beta$  and its prior mean. We find that we can get the priors for  $p_s$  and  $\gamma$  from the priors for  $\alpha$  and  $\beta$  by the absolute Jacobian of the transformation, however, it may be difficult or impossible to obtain the

priors for  $\alpha$  and  $\beta$  from the priors for  $p_s$  and  $\gamma$  because we need to know  $\alpha_0$  and  $\beta_0$  in advance but they are related with the priors for  $\alpha$  and  $\beta$ .

We will now look at the corresponding prior distributions on  $p_s$  when  $\alpha$  takes different fixed values and different prior distributions described in Table 5.1. We assume the prior distribution of  $\beta$  is normal with mean 0 and variance 1.34, which is the default prior distribution for  $\beta$  in the R package “dfcrm”. We also assume the starting dose level is the lowest dose level, and the initial guess for its toxicity probability is 0.05.

Table 5.1 Different Choices of Values/Prior Distributions for  $\alpha$

$\alpha$	Value/Prior Distribution
	-6
	-3
	0
	3
	6
	Normal with mean 3 and variance 1
	Normal with mean -6 and variance 10
	Normal with mean 3 and variance 100

To obtain the prior density function for  $p_s$ , we create a set of 10000 random draws from the prior distribution for  $\beta$ , and also for  $\alpha$  if  $\alpha$  is not a fixed parameter, then calculate  $p_s$  by applying Equation (14) each of the 10000 parameter pairs  $(\alpha_i, \beta_i)$ . For each scenario, we draw the prior density plot and the empirical cumulative distribution function (CDF) plot for  $p_s$  based on the simulated data.

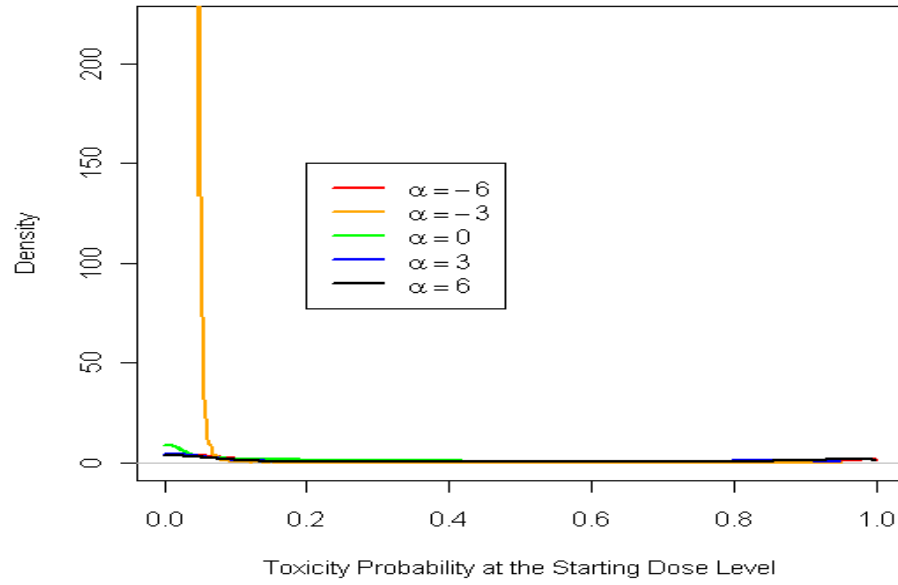


Figure 5.2 Prior Density Plot of  $p_s$  when  $\alpha$  is Fixed

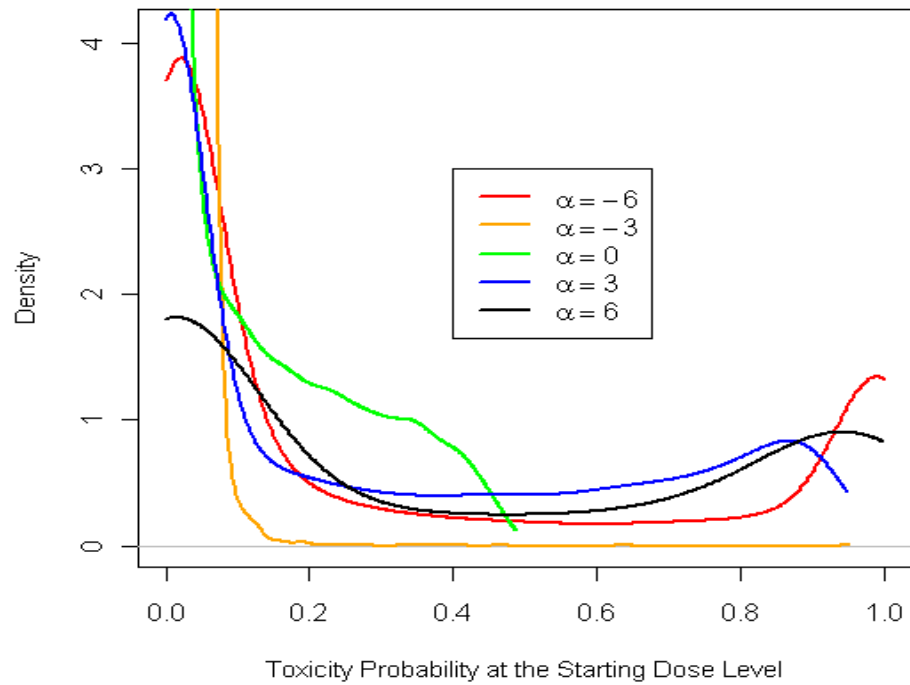


Figure 5.3 Enlarged Prior Density Plot of  $p_s$  when  $\alpha$  is Fixed

Figure 5.2 and Figure 5.3 display the prior density plot of  $p_s$  when  $\alpha$  takes different fixed values. The initial guessed log odds of having toxicity response,  $\log(\frac{\pi_s}{1-\pi_s})$ , is equal to -2.94, which is very close to -3. When  $\alpha$  takes -3,  $\log(\frac{\pi_s}{1-\pi_s}) - \alpha_0$  in the Equation (14) is close to 0 such that  $p_s$  depends very little on  $\beta$ , which is the only source of the variation. As we can see from the above plots, if we choose -3 for  $\alpha$ , the prior density of  $p_s$  is close to point mass at 0.05, which would be only appropriate if, at the beginning of a trial, the investigator is nearly 100% sure that the toxicity probability at the starting dose level is 0.05. From the original parameterization this is not at all obvious. When  $\alpha$  takes fixed values -6, 3 and 6, we observe that the corresponding prior distribution of  $p_s$  is bimodal with a mode near 0 and a mode near 1. The reason for this is because of Jacobian for transforming the probability from the logit scale to the original scale. Although the left-hand mode is higher than the right-hand mode, this bimodality raises the worry whether a design would behave unexpectedly if the first few accruals experience toxicity responses, to be discussed in Section 5.1.3.

Figure 5.4 displays the empirical cumulative distribution (ECDF) of  $p_s$  calculated from the simulated data. The ECDF curve is very steep when  $\alpha = -3$ , and the other curves except for the one when  $\alpha = 0$  show that we initially assume there are more than 20% chances that the toxicity probability at the starting dose level is larger than 0.6. In Phase I clinical trials, it would seem likely that investigators would usually be uncomfortable agreeing with this belief. Therefore, assuming fixed values of -6, 3 and 6 for  $\alpha$  seems not quite realistic under our setting for  $\pi_s$  and prior distribution of  $\beta$ .



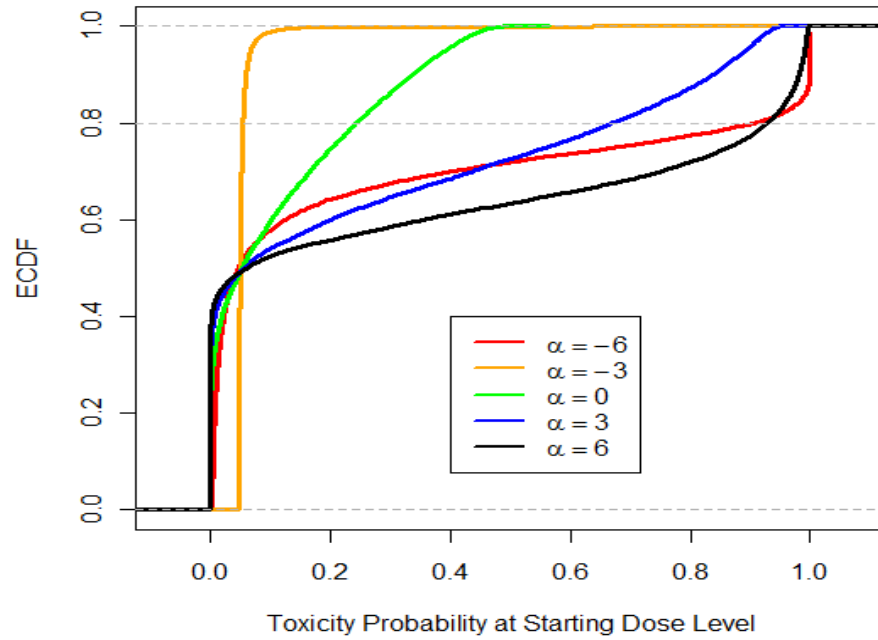


Figure 5.4 ECDF Plot of  $p_s$  when  $\alpha$  is Fixed

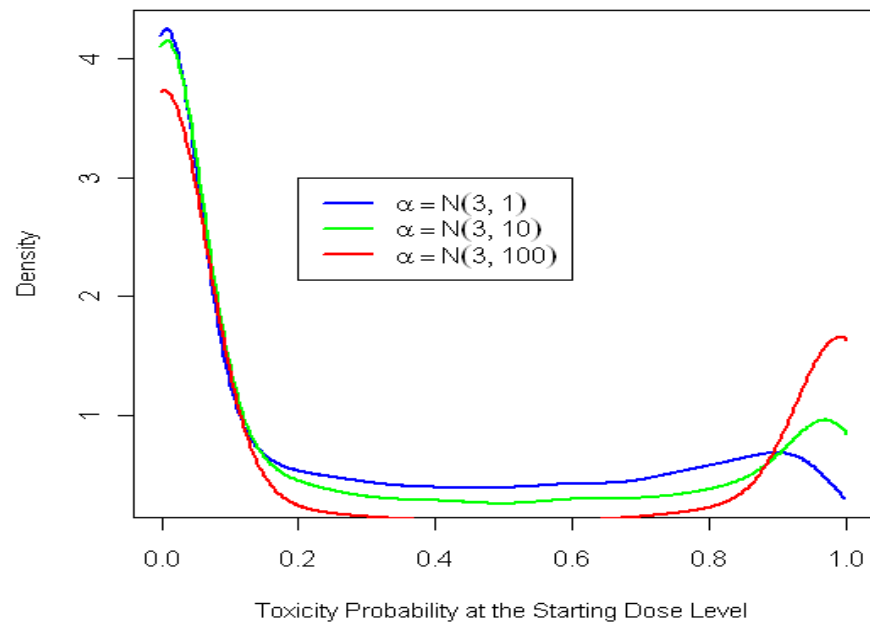


Figure 5.5 Prior Density Plot of  $p_s$  when  $\alpha$  is Free

Figure 5.5 displays the prior density plot of  $p_s$  when  $\alpha$  follows normal distribution with mean 3 and different variances. All the corresponding prior distributions for  $p_s$  are bimodal at two ends. An interesting finding is that as the variance increases, the left-hand mode lowers, while the right-hand mode rises.

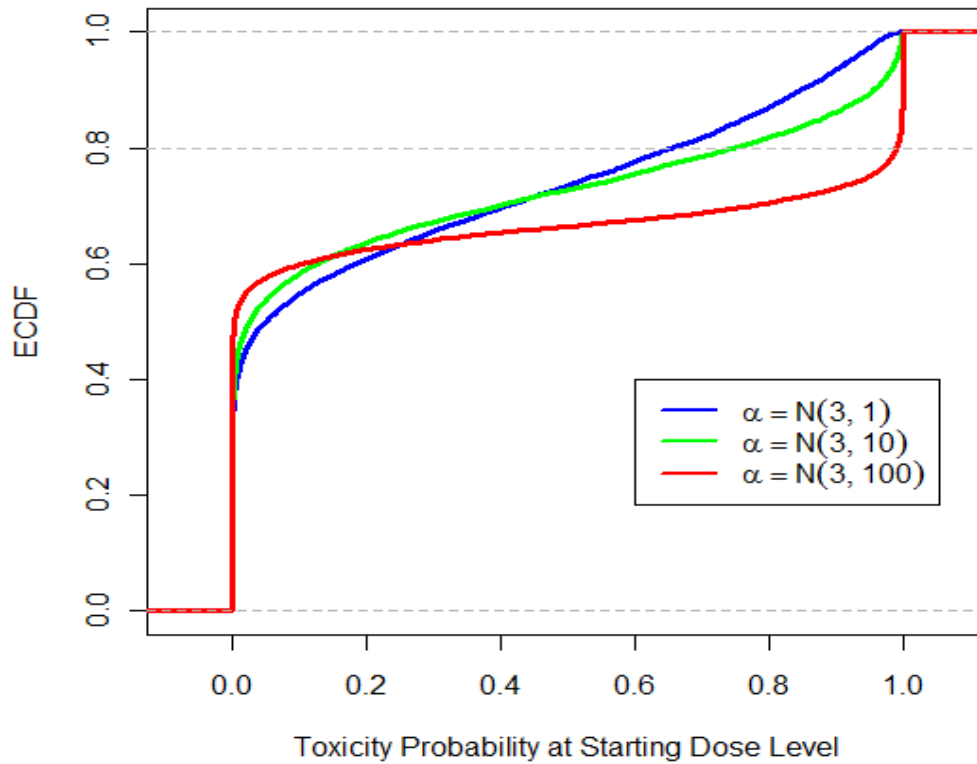


Figure 5.6 ECDF Plot of  $p_s$  when  $\alpha$  is Free

Figure 5.6 displays the empirical cumulative distribution of  $p_s$  calculated from the simulated data when  $\alpha$  is free. For each of these priors, the chance exceeds 20% that the toxicity risk at the starting dose level is larger than 0.6. For the same reason illustrated in the previous paragraph, the prior distributions for  $p_s$  induced from those prior distributions for  $\alpha$  are not sensible in most of Phase I clinical trials.

In summary, the lessons we conclude from the above investigation are:

- The choices of fixed value or prior distributions for  $\alpha$  in the conventional logit model are not arbitrary. We should consider carefully about their implications for the prior distribution for  $p_s$ . The typical fixed value of 3 that  $\alpha$  takes may not be reasonable under some settings.
- We recommend using the re-parameterized logit model with interpretable parameters, which can facilitate the elicitation of sensible priors that genuinely reflect investigators' prior belief.

### **5.1.3 Choices of prior distribution for $p_s$**

In this section, we will investigate how different choices of prior distribution for  $p_s$  affect CRM design with respect to the dose assignments for the first five patients, the average number of toxicity responses, the average proportion of toxicity responses, the average percentage of patients treated at the true target dose level, proportion of simulations where the true target dose level is estimated correctly and the proportion of early stopping.

#### **5.1.3.1 Experiment set-up**

##### **Designs:**

All the testing CRM designs share the following common design parameters:

- Number of testing dose levels: 5
- The starting dose level is the first dose level
- Single-stage

- One patient per cohort
- Maximum sample size: 30
- Target toxicity rate: 0.25
- Early Stopping: the posterior toxicity probability at the first dose level is larger than  $(0.1 + \text{target}) = 0.35$
- $\gamma$ : fixed at 1
- Initial toxicity probability guesses:

$$\text{logit}(p_j) = -3.86 + 0.92j \quad (16)$$

, where  $\text{logit}(p_j) = \log\left(\frac{p_j}{1-p_j}\right)$ , and  $j$  refers to the dose level  $j$ .

Equation (16) shows the initially guessed outcome model, and its corresponding dose-toxicity response curve is shown in Figure 5.7, and Table 5.2 lists the initially guessed toxicity probabilities at the testing dose levels.

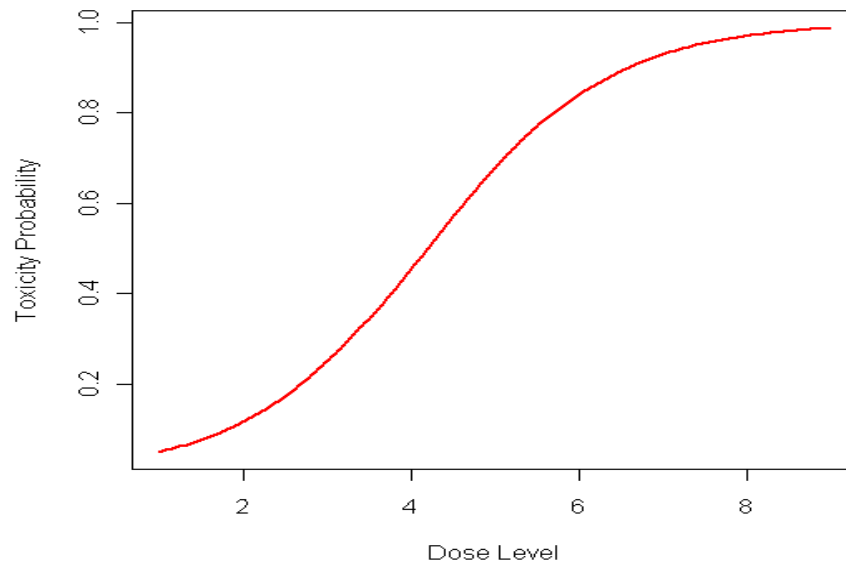


Figure 5.7 Initially Guessed Dose-toxicity Response Curve

Table 5.2 Initially Guessed Toxicity Probabilities

	Dose Level				
	1	2	3	4	5
<b>Pr(Toxicity)</b>	0.05	0.12	0.25	0.46	0.68

Table 5.3 CRM Designs with Different Priors for  $p_s$ 

Design #	Prior	Dose Escalation Restriction	Design Name
1	$p_s \sim \text{Beta}(1.47, 10); \gamma = 1$	No	“Informative”
2	$p_s \sim \text{Beta}(1.11, 3); \gamma = 1$	No	“Mild”
3	$\alpha = 3, \beta = N(0, 1.34)$	No	“Bimodal”
4	$p_s \sim \text{Uniform}(0, 1); \gamma = 1$	No	“Uniform”
5	$p_s \sim \text{Beta}(1.47, 10); \gamma = 1$	Yes	“Informative, restriction”
6	$p_s \sim \text{Beta}(1.11, 3); \gamma = 1$	Yes	“Mild, restriction”
7	$\alpha = 3, \beta = N(0, 1.34)$	Yes	“Bimodal, restriction”
8	$p_s \sim \text{Uniform}(0, 1); \gamma = 1$	Yes	“Uniform, restriction”

Designs differ only with respect to the priors for  $p_s$  and the use of dose escalation restriction. For comparison, we also include the CRM design when  $\alpha$  is 3 and the prior for  $\beta$  follows normal distribution with mean 0 and variance 1.34. In our current implementation of the platform, dose escalation restriction includes no skipping doses and no dose escalation immediately after a toxicity response. Table 5.3 lists all the testing designs along with their distinguishing features and names here. The prior Beta(1.47,10) assumes  $p_s$  falls within the interval (0,0.1) with the probability of 0.47; the prior Beta(1.11,3) assumes with the probability

of 0.23, the induced prior from  $\alpha=3$  and  $\beta=N(0,1.34)$  assumes with the probability of 0.54, and the prior uniform (0,1) assumes with the probability of 0.1. The prior density plots for  $p_s$  are drawn in Figure 5.8.

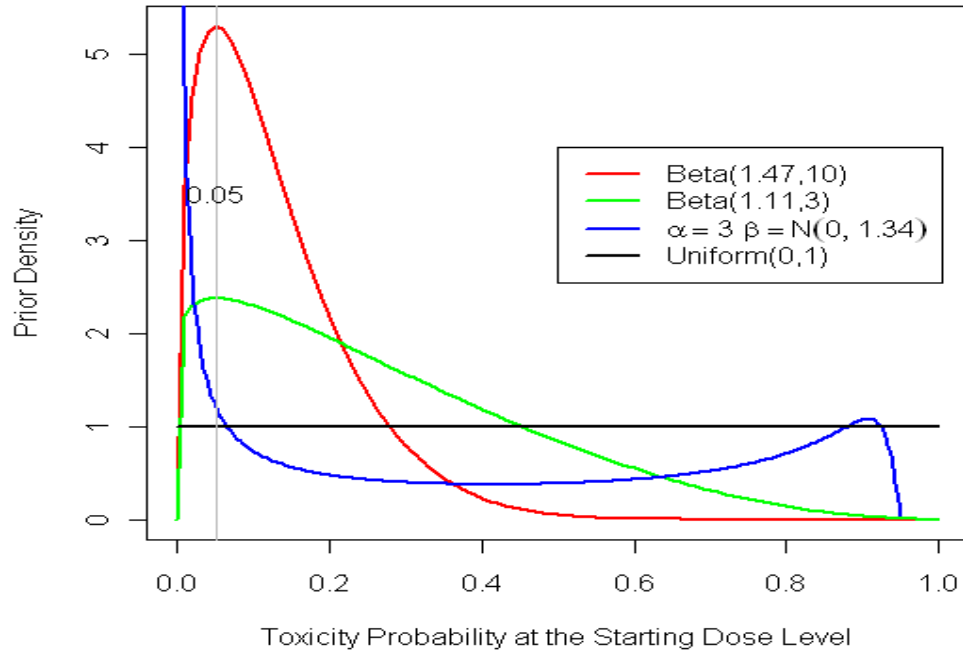


Figure 5.8 Prior Density Plots of  $p_s$

Table 5.4 Toxicity Probabilities from the Dose-toxicity Curves with  $\gamma=1$

		Tier Dose				
		1	2	3	4	5
Model # (Name)	1 ("Guess=Truth")	0.05	0.12	0.25	0.46	0.68
	2 ("Guess<Truth")	0.25	0.46	0.68	0.84	0.93
	3 ("Guess>Truth")	0.0082	0.02	0.049	0.12	0.25

**Population Model:** None, which assumes no baseline characteristics affect either the decision-making in the designs or patients’ outcomes.

### Outcome Models:

We consider three underlying outcome models which correspond to the dose-toxicity curves with the same shape as the initially guessed ( $\gamma=1$ ), as shown in Figure 5.9. The outcome model corresponding to the red curve assumes all  $\pi_j = p_j$  and we denote it as “Guess=Truth”; the outcome model corresponding to the black solid curve assumes all  $\pi_j < p_j$  and we denote it as “Guess<Truth”; the outcome model corresponding to the black dashed curve assumes all  $\pi_j > p_j$  and we denote it as “Guess>Truth”. The underlying toxicity probabilities at the testing dose levels are listed in Table 5.4. The true target dose level is 3 in the model “Guess=Truth”, 1 in the model “Guess<Truth”, and 3 in the model “Guess>Truth”.

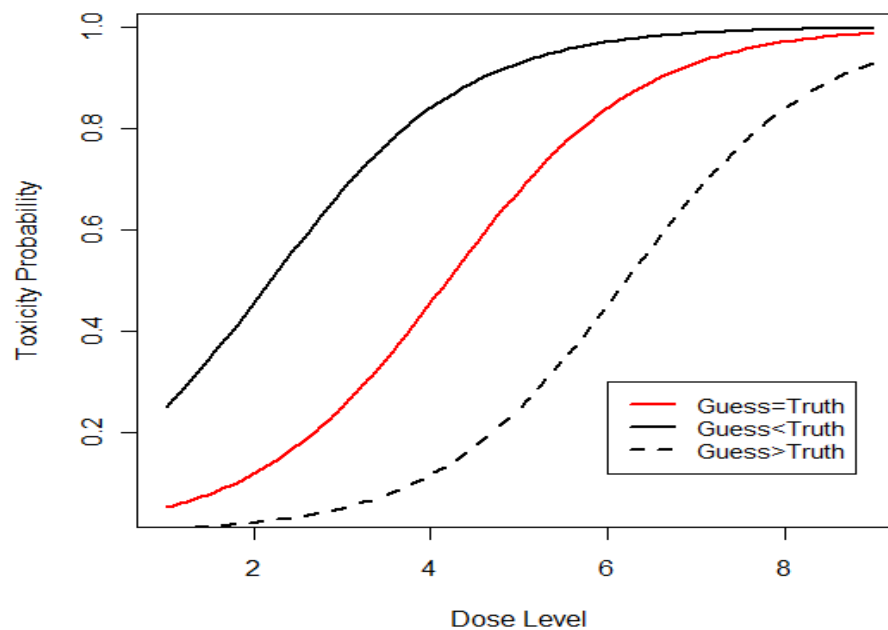


Figure 5.9 Underlying Dose-toxicity Curves with  $\gamma=1$

**Evaluation Criteria:**

- Average number of toxicity responses
- Average proportion of toxicity responses
- Average percentage of patients treated at the true target dose level
- Proportion of simulations where the true target dose level is estimated correctly
- Proportion of early stopping

**Number of Replications: 500**

This experiment includes  $8 \times 3 = 24$  scenarios for simulation.

**5.1.3.2 Results**

The accumulating data at the beginning of a clinical trial are very sparse, thus the early decisions in a CRM design are strongly influenced by the assumed prior distributions. For this reason, appropriate choice of prior distributions is very important to ensure ethical and efficient early decision makings. Next, we look at how different priors for  $p_s$  affect the dose assignments for the first five patients.

Figure 5.10 - Figure 5.13 display the dose assignments for the first five patients in the four testing designs without the dose escalation restriction respectively. Each arrow is away from assigned dose level for the previous patient, and points to the assigned dose level for the following patient. Blue arrows indicate previous patients do not have toxicity response after receiving the assigned dose, and red arrows indicate previous patients have toxicity response. The number in each rectangular box is the assigned dose level. For example, in the “informative” design, the first patient is assigned the first dose level, if he/she does not have a toxicity



response, the next patient will be assigned the second dose level, otherwise, the next patient will be assigned the first dose level.

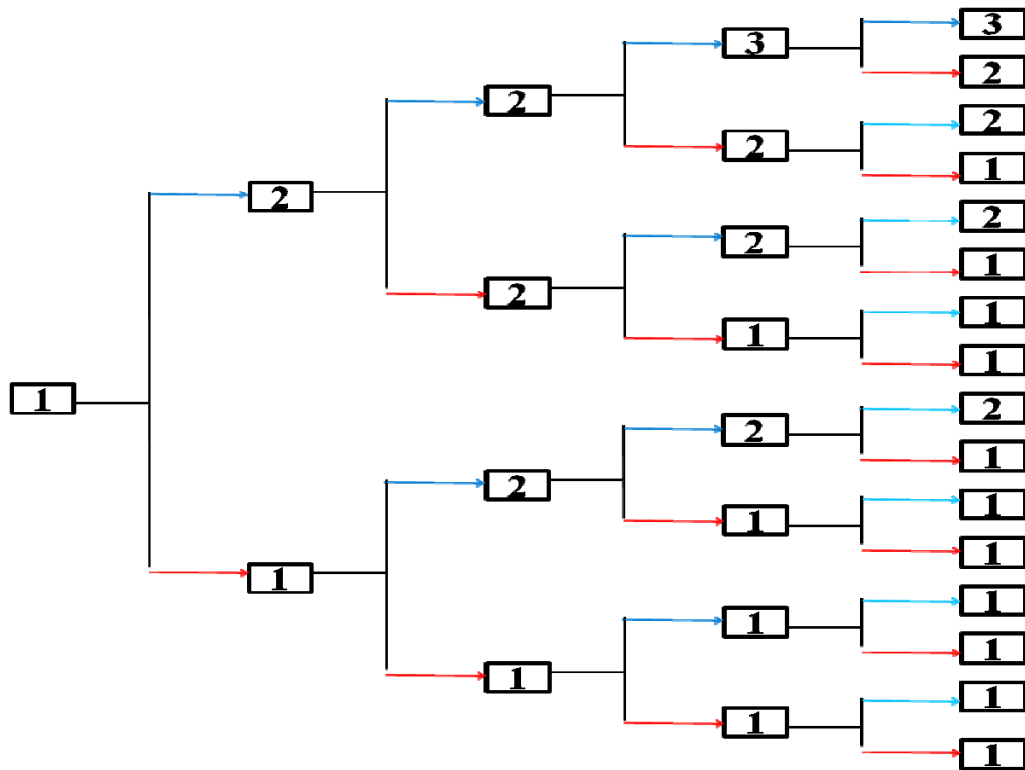


Figure 5.10 Dose Assignments for the First Five Patients in the “Informative” Design

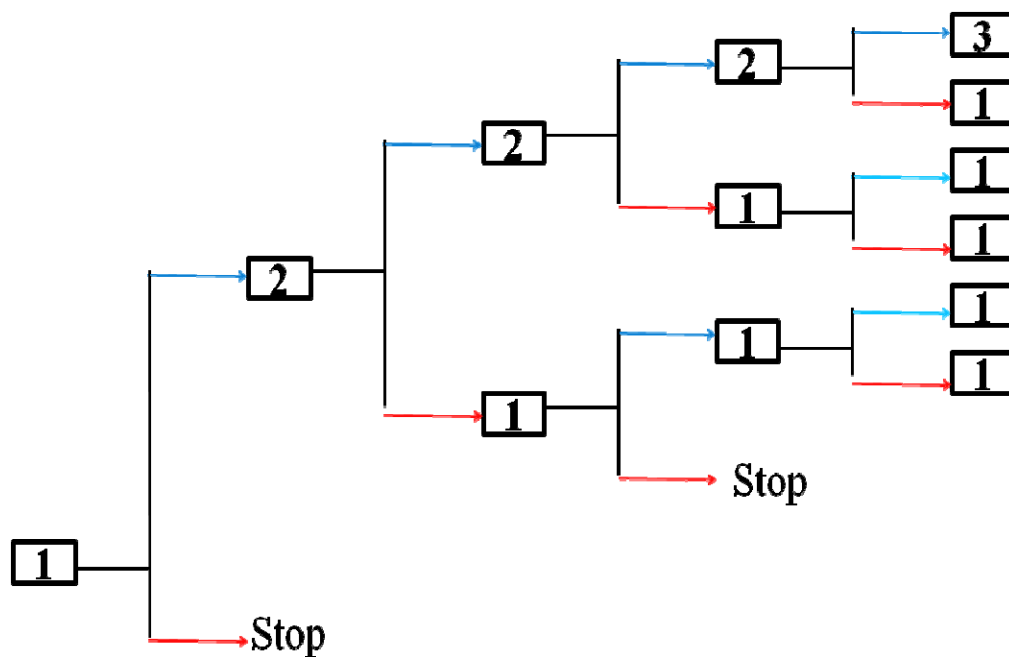


Figure 5.11 Dose Assignments for the First Five Patients in the “Mild” Design

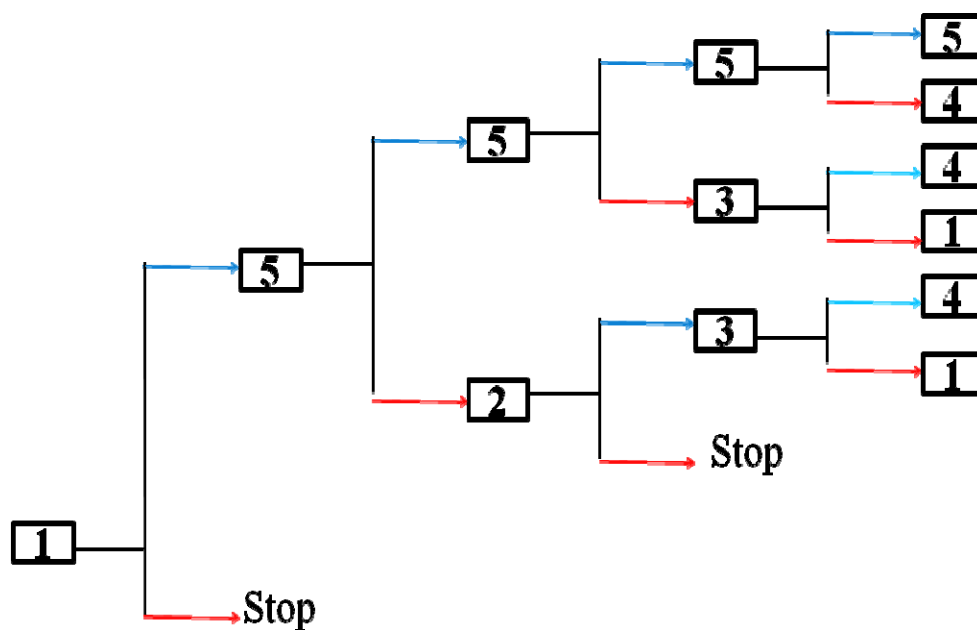


Figure 5.12 Dose Assignments for the First Five Patients in the “Bimodal” Design

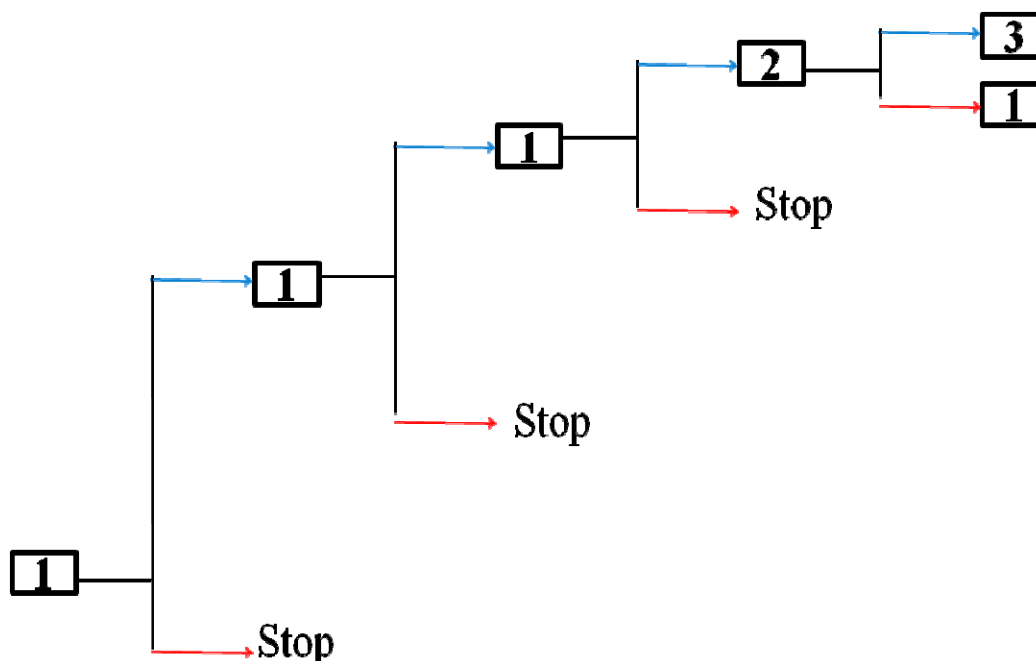


Figure 5.13 Dose Assignments of the First Five Patients in the “Uniform” Design

From the above four figures, we observe that in the “informative” design no early stopping occurs even if the first four enrolled patients all have toxicities. In the other three designs, clinical trials are stopped early when even one toxicity response is seen at low doses. Compared to the “informative” design, the “mild” design is more conservative, and it generally either assigns a lower dose level to the next patient or stops trials given the same accumulating data. The “Bimodal” design is the most aggressive when data are sparse, and it assigns the highest dose level to the second patient when the first patient does not have a toxicity response, which is probably because the mode of the posterior distribution quickly shifts to the lower end after observing an non-toxicity response. We think this might be the reason why the CRM has been criticized for being too aggressive in dose escalation and has been recommended to be used

with overdose control [41]. The “uniform” design is the most conservative with very slow escalation.

In general, incorporating the dose escalation restriction into a design makes dose assignment more conservative. For the first five patients’ dose assignments, this effect is only seen in the “Bimodal” design where skipping dose levels occur. Figure 5.14 shows the dose assignments for the first five patients in the “Bimodal, restriction” design.

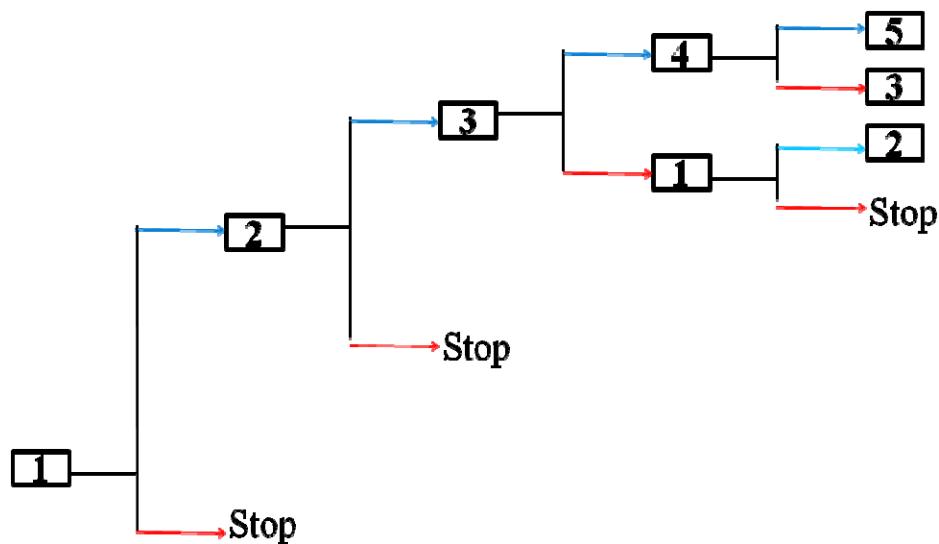


Figure 5.14 Dose Assignments for the First Five Patients in the “Bimodal, restriction” Design

Furthermore, if we know the underlying toxicity probabilities at each dose level, we can easily calculate the probability of the  $n$ th ( $n \leq 5$ ) patient being assigned a certain dose level from the multiplication of the probabilities before this patient. For example, the probability of the third patient being assigned the 3rd dose level in the “Informative” design with the underlying “Guess=Truth” outcome model would be:  $0.95 \times 0.88 \times 0.88 = 0.74$ .

We will look at the evaluation results based on different specified criteria.

- Average number of toxicity responses

Table 5.5 Number of Toxicity Responses when “Guess=Truth”

		Number of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>	“Informative”	6.0	0.063
	“Mild”	5.8	0.077
	“Bimodal”	7.4	0.12
	“Uniform”	4.8	0.092
	“Informative, restriction”	6.1	0.061
	“Mild, restriction”	5.5	0.077
	“Bimodal, restriction”	6.8	0.129
	“Uniform, restriction”	4.9	0.095

Table 5.6 Number of Toxicity Responses when “Guess<Truth”

		Number of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>	“Informative”	9.5	0.073
	“Mild”	5.5	0.15
	“Bimodal”	3.9	0.16
	“Uniform”	3.3	0.14
	“Informative, restriction”	9.5	0.071
	“Mild, restriction”	5.2	0.155
	“Bimodal, restriction”	3.1	0.147
	“Uniform, restriction”	3.3	0.144

Table 5.5 - Table 5.7 list the average numbers of toxicity responses and their standard errors across different designs and different underlying outcome models. Designs assuming the bimodal prior have the largest standard errors among the designs under the same underlying outcome models. The standard errors increase with the increase of the aggressiveness of the initial guesses. Incorporating the dose escalation restriction into a design is found not to change

the average number of toxicity responses very much. The average number of toxicity responses is the largest in the designs assuming the bimodal prior when “Guess=Truth” or “Guess>Truth”, and it is the largest in the designs assuming the informative prior when “Guess<Truth”. We observe that assuming the uniform prior will decrease the average number of toxicity responses in most cases compared to assuming the other priors.

Table 5.7 Number of Toxicity Responses when “Guess>Truth”

		Number of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>	“Informative”	3.7	0.054
	“Mild”	3.8	0.057
	“Bimodal”	5.9	0.079
	“Uniform”	3.6	0.056
	“Informative, restriction”	3.7	0.053
	“Mild, restriction”	3.9	0.059
	“Bimodal, restriction”	5.3	0.084
	“Uniform, restriction”	3.6	0.055

- Average proportion of toxicity responses

Table 5.8 - Table 5.10 list the average proportions of toxicity responses and their standard errors across different designs and different underlying outcome models. The standard errors increase with the increase of the aggressiveness of the initial guesses. Incorporating the dose escalation restriction into a design is found not to change the average number of toxicity responses very much, except the designs assuming the bimodal prior when “Guess<Truth”. Regardless of the underlying outcome models, the designs assuming the informative prior have the smallest average proportion of toxicity responses.

Table 5.8 Proportion of Toxicity Responses when “Guess=Truth”

		Proportion of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>			
	“Informative”	0.201	0.002
	“Mild”	0.229	0.007
	“Bimodal”	0.346	0.008
	“Uniform”	0.252	0.009
	“Informative, restriction”	0.203	0.002
	“Mild, restriction”	0.231	0.008
	“Bimodal, restriction”	0.328	0.008
	“Uniform, restriction”	0.255	0.009

Table 5.9 Proportion of Toxicity Responses when “Guess<Truth”

		Proportion of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>			
	“Informative”	0.332	0.004
	“Mild”	0.509	0.013
	“Bimodal”	0.665	0.01
	“Uniform”	0.529	0.013
	“Informative, restriction”	0.332	0.004
	“Mild, restriction”	0.535	0.013
	“Bimodal, restriction”	0.576	0.011
	“Uniform, restriction”	0.515	0.013

Table 5.10 Proportion of Toxicity Responses when “Guess>Truth”

		Proportion of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>			
	“Informative”	0.124	0.002
	“Mild”	0.137	0.004
	“Bimodal”	0.201	0.003
	“Uniform”	0.132	0.004
	“Informative, restriction”	0.123	0.002
	“Mild, restriction”	0.135	0.004
	“Bimodal, restriction”	0.194	0.004
	“Uniform, restriction”	0.129	0.003

- Average percentage of patients treated at the true target dose level

Table 5.11 Percentage of Patients Treated when “Guess=Truth”

		<b>Below Target</b>	<b>At Target</b>	<b>Above Target</b>
<b>Design</b>	“Informative”	46.2	47.4	6.4
	“Mild”	50.6	42.4	7.1
	“Bimodal”	31.6	42.7	25.7
	“Uniform”	58.5	35.1	6.4
	“Informative, restriction”	44.0	48.6	7.5
	“Mild, restriction”	52.6	40.3	7.1
	“Bimodal, restriction”	36.5	44.1	19.4
	“Uniform, restriction”	58.7	33.9	7.3

Table 5.12 Percentage of Patients Treated when “Guess<Truth”

		<b>Below Target</b>	<b>At Target</b>	<b>Above Target</b>
<b>Design</b>	“Informative”	0	66.6	33.4
	“Mild”	0	78.6	21.4
	“Bimodal”	0	55.8	44.2
	“Uniform”	0	89.4	10.6
	“Informative, restriction”	0	66.5	33.5
	“Mild, restriction”	0	80.4	19.6
	“Bimodal, restriction”	0	62.0	38.0
	“Uniform, restriction”	0	89.5	10.5

Table 5.11 - Table 5.13 list the average percentages of patients treated below, at and above the true target dose level across different designs and different underlying outcome models. The true target dose level is 3 when “Guess=Truth”, 1 when “Guess<Truth”, and 5 when “Guess>Truth”. Incorporating the dose escalation restriction into a design is found not to change the average percentage of patients very much except for the designs assuming the bimodal prior. When “Guess=Truth”, the average percentage of patients treated at the true target dose level is



the largest in the designs assuming the informative prior, the “Mild” design and the “Bimodal” design have the very similar percentage at the target, however quite different percentages below and above the target. The average percentage at the target is the largest in the designs assuming the uniform prior when “Guess<Truth” and in the designs assuming the bimodal prior when “Guess>Truth”.

Table 5.13 Percentage of Patients Treated when “Guess>Truth”

		<b>Below Target</b>	<b>At Target</b>	<b>Above Target</b>
<b>Design</b>	“Informative”	72.4	27.6	0
	“Mild”	69.6	30.4	0
	“Bimodal”	30.3	69.7	0
	“Uniform”	69.2	30.8	0
	“Informative, restriction”	76.7	23.3	0
	“Mild, restriction”	68.0	32.0	0
	“Bimodal, restriction”	42.0	58.0	0
	“Uniform, restriction”	70.6	29.4	0

- Proportion of simulations where the true target dose level is estimated correctly

Table 5.14 Proportion of Correct Estimation of the True Target when “Guess=Truth”

		<b>Correct Estimation of the True Target</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Informative”	0.616	0.572	0.659
	“Mild”	0.602	0.558	0.645
	“Bimodal”	0.612	0.568	0.655
	“Uniform”	0.508	0.463	0.553
	“Informative, restriction”	0.634	0.59	0.676
	“Mild, restriction”	0.582	0.537	0.626
	“Bimodal, restriction”	0.63	0.586	0.672
	“Uniform, restriction”	0.51	0.465	0.555

Table 5.15 Proportion of Correct Estimation of the True Target when “Guess&lt;Truth”

		<b>Correct Estimation of the True Target</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Informative”	0.818	0.781	0.851
	“Mild”	0.448	0.404	0.493
	“Bimodal”	0.14	0.111	0.174
	“Uniform”	0.234	0.198	0.274
	“Informative, restriction”	0.824	0.788	0.856
	“Mild, restriction”	0.418	0.374	0.463
	“Bimodal, restriction”	0.136	0.107	0.169
	“Uniform, restriction”	0.204	0.17	0.242

Table 5.16 Proportion of Correct Estimation of the True Target when “Guess&gt;Truth”

		<b>Correct Estimation of the True Target</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Informative”	0.394	0.351	0.438
	“Mild”	0.418	0.374	0.463
	“Bimodal”	0.778	0.739	0.814
	“Uniform”	0.498	0.453	0.543
	“Informative, restriction”	0.352	0.31	0.396
	“Mild, restriction”	0.398	0.355	0.442
	“Bimodal, restriction”	0.724	0.683	0.763
	“Uniform, restriction”	0.478	0.433	0.523

Table 5.14 - Table 5.16 list the proportions of correct estimation of the true target and their exact 95% confidence intervals across different designs and different underlying outcome models. Incorporating the dose escalation restriction into a design is found not to change the proportion of correct estimation of the true target very much except that it decreases the proportion significantly for the design assuming bimodal prior when “Guess>Truth”. When “Guess=Truth”, the design assuming uniform prior has the smallest proportion, while the designs

assuming other kinds of priors have similar proportions; when “Guess<Truth”, designs assuming informative prior has significant larger proportion than the other designs, while designs assuming bimodal priors has the smallest proportion; when “Guess>Truth”, designs assuming bimodal prior has the largest proportion, designs assuming uniform prior has the second largest, and designs assuming informative and mild priors have similar proportions.

- Proportion of early stopping

Table 5.17 - Table 5.19 list the proportions of early stopping and their exact 95% confidence intervals across different designs and different underlying outcome models. Under the settings in this experiment, since the true target dose level is within the testing dose levels, early stopping is not favorable. Incorporating the dose escalation restriction into a design is found to increase a little bit the proportion of early stopping except the designs assuming the mild or uniform prior when “Guess>Truth”. Regardless of the underlying outcome models, the designs assuming the informative prior have the smallest proportion of early stopping.

Table 5.17 Proportion of Early Stopping when “Guess=Truth”

		<b>Early Stopping</b>		
		Proportion	95% CI	
			Lower Bound	Upper Bound
<b>Design</b>	“Informative”	0	0	0.006
	“Mild”	0.042	0.026	0.063
	“Bimodal”	0.14	0.111	0.174
	“Uniform”	0.154	0.123	0.189
	“Informative, restriction”	0	0	0.006
	“Mild, restriction”	0.052	0.034	0.075
	“Bimodal, restriction”	0.164	0.133	0.199
	“Uniform, restriction”	0.164	0.133	0.199

Table 5.18 Proportion of Early Stopping when “Guess&lt;Truth”

		<b>Early Stopping</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Informative”	0.044	0.028	0.066
	“Mild”	0.492	0.447	0.537
	“Bimodal”	0.814	0.777	0.847
	“Uniform”	0.714	0.672	0.753
	“Informative, restriction”	0.046	0.029	0.068
	“Mild, restriction”	0.526	0.481	0.57
	“Bimodal, restriction”	0.83	0.794	0.862
	“Uniform, restriction”	0.74	0.699	0.778

Table 5.19 Proportion of Early Stopping when “Guess&gt;Truth”

		<b>Early Stopping</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Informative”	0	0	0.006
	“Mild”	0.012	0.004	0.026
	“Bimodal”	0.006	0.001	0.017
	“Uniform”	0.016	0.007	0.031
	“Informative, restriction”	0	0	0.006
	“Mild, restriction”	0.006	0.001	0.017
	“Bimodal, restriction”	0.032	0.018	0.051
	“Uniform, restriction”	0.014	0.006	0.029

### 5.1.3.3 Conclusions

Results in Section 5.1.3.2 show that no designs perform uniformly best across different criteria and different underlying outcome models. A design may perform best according to a certain criterion regardless of the underlying outcome models, for example, the designs assuming the

informative prior (with and without dose escalation restriction) have the smallest proportion of early stopping. Incorporating dose escalation restriction to the designs assuming the informative prior, mild prior, or uniform prior is found not to affect the design performance very much. The design assuming the bimodal prior, however, benefits significantly from the overdose control except for the case when the underlying outcome model is “Guess>Truth” and the evaluation criteria are the average percentage of patients treated at the true target dose level and the proportion of simulations where the true target dose level is estimated correctly.

#### **5.1.4 Choices of prior distribution for $\gamma$**

In this section, we will investigate how different choices of prior distribution for  $\gamma$  affect CRM design with respect to the dose assignments for the first five patients, the average number of toxicity responses, the average proportion of toxicity responses, the average percentage of patients treated at the true target dose level, proportion of simulations where the true target dose level is estimated correctly and the proportion of early stopping.

##### **5.1.4.1 Experiment set-up**

###### **Designs:**

All the testing CRM designs share the following common design parameters:

- Number of testing dose levels: 5
- The starting dose level is the first dose level
- Single-stage
- One patient per cohort
- Maximum sample size: 30

- Target toxicity rate: 0.25
- Early Stopping: the posterior toxicity probability at the first dose level is larger than  $(0.1 + \text{target}) = 0.35$
- Prior for  $p_s$ : Beta(1.47,10)
- Initial toxicity probability guesses: the same as those in the experiment described in Section 5.1.3.1.
- No dose escalation restriction

Designs differ only with respect to the priors for  $\gamma$ . Figure 5.15 displays the three prior distributions under investigation. The prior LogN(0, 0.04) assumes that  $\log(\gamma)$  has normal prior with mean 0 and variance 0.04 and that the probability of either  $\gamma \leq 0.5$  or  $\gamma \geq 2$  is almost 0, the prior LogN(0, 2.89) assumes  $\log(\gamma)$  has normal prior with mean 0 and variance 2.89 and that the probability of either  $\gamma \leq 0.5$  or  $\gamma \geq 2$  is 0.41, and the prior 1 assumes  $\gamma$  is fixed at 1. The priors LogN(0,0.04) and LogN(0,2.89) both have their medians at 1. To help remembering different designs' features, we name them according to their priors for  $\gamma$ . Table 5.20 lists the design names along with the corresponding prior assumptions.

Table 5.20 CRM Designs with Different Priors for  $\gamma$

Design #	Prior	Design Name
1	$p_s \sim \text{Beta}(1.47, 10); \gamma \sim \text{LogN}(0, 0.04)$	"Small variance"
2	$p_s \sim \text{Beta}(1.47, 10); \gamma \sim \text{LogN}(0, 2.89)$	"Big variance"
3	$p_s \sim \text{Beta}(1.47, 10); \gamma = 1$	"Fixed"

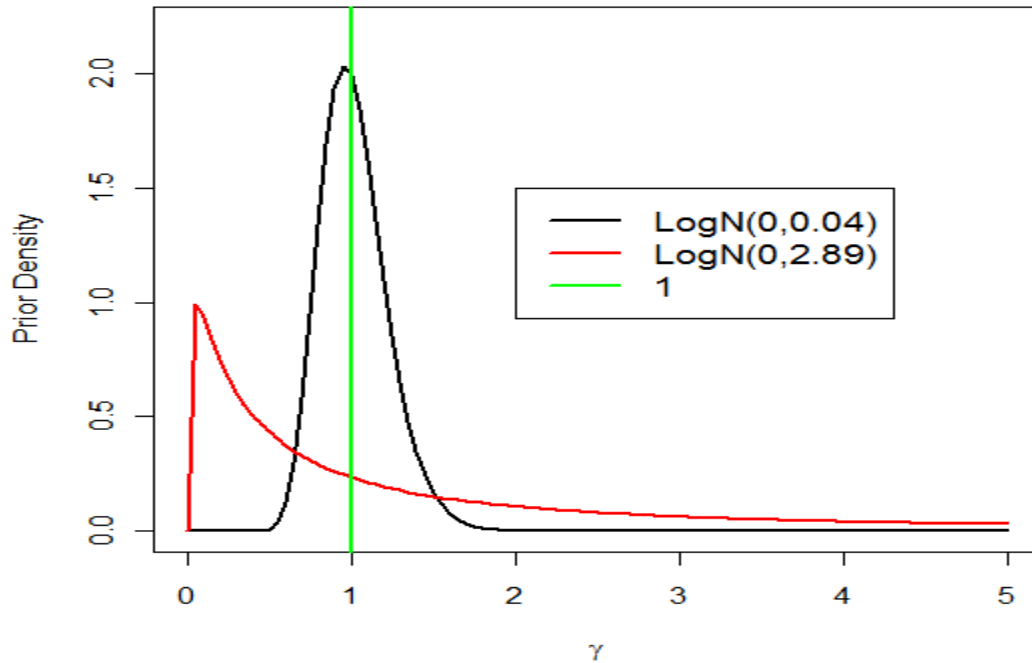


Figure 5.15 Prior Density Plots of  $\gamma$

**Population Model:** None, which assumes no baseline characteristics affect either the decision-making in the designs or patients' outcomes.

**Outcome Models:**

Table 5.21 Toxicity Probabilities from the Dose-toxicity curves with Different  $\gamma$

		Tier Dose				
		1	2	3	4	5
Model # (Name)	1 (" $\gamma = 1$ ")	0.05	0.12	0.25	0.46	0.68
	2 (" $\gamma = 2$ ")	0.05	0.25	0.67	0.93	0.99
	3 (" $\gamma = 0.5$ ")	0.05	0.08	0.12	0.17	0.25

We consider three underlying outcome models. Two of them correspond to the dose-toxicity curves with different shapes from the initially guessed, and one is the same as the initially guessed. The curves are shown in Figure 5.16. The underlying toxicity probabilities at the testing dose levels are listed in Table 5.21. The true target dose level is 3 in the model “ $\gamma = 1$ ”, 2 in the model “ $\gamma = 2$ ”, and 3 in the model “ $\gamma = 0.5$ ”.

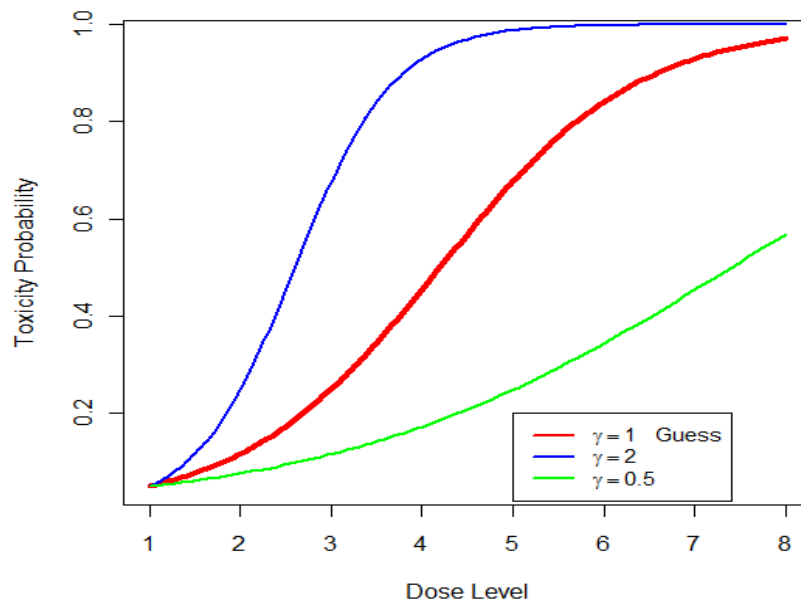


Figure 5.16 Underlying Dose-toxicity Curves with Different  $\gamma$

#### Evaluation Criteria:

- Average number of toxicity responses
- Average proportion of toxicity responses
- Average percentage of patients treated at the true target dose level
- Proportion of simulations where the true target dose level is estimated correctly
- Proportion of early stopping



**Number of Replications: 500**

This experiment includes  $3 \times 3 = 9$  scenarios for simulation.

### 5.1.4.2 Results

First, we look at how different priors for  $\gamma$  affect the dose assignments for the first five patients.

The “Small variance” design and “Fixed” designs have the same dose assignments for the first five patients, which is shown in Figure 5.17. Figure 5.18 displays the dose assignments for the first five patients in the “Big variance” design. We observe that no early stopping occurs before the fifth enrolled patients in these three designs. The “Big variance” design escalates and de-escalates dose levels more aggressively than the other two designs.

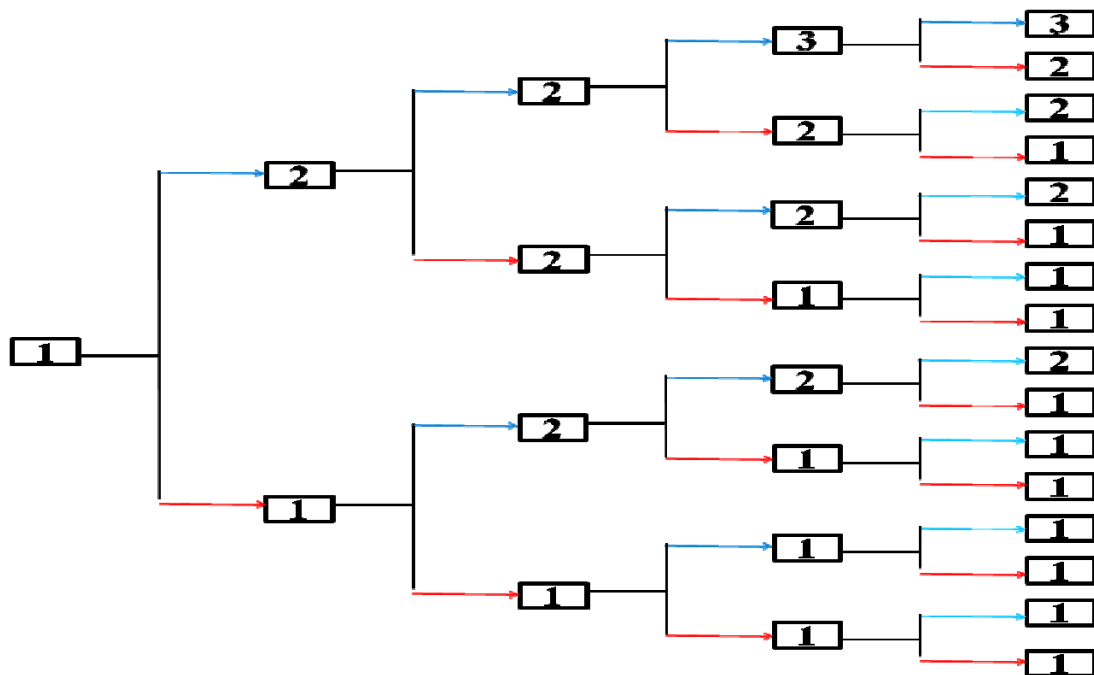


Figure 5.17 Dose Assignments for the First Five Patients in the “Small variance” or “Fixed”

## Design



increase with the increase of the variance of the prior for  $\gamma$ . The average number of toxicity responses is the smallest in the “Fixed” design when “ $\gamma=1$ ” or “ $\gamma=0.5$ ”, and in the “Big variance” design when “ $\gamma=2$ ”. Adding a small randomness to the fixed prior seems not to improve the design performance with respect to the average number of toxicity responses.

Table 5.23 Number of Toxicity Responses when “ $\gamma=2$ ”

		Number of Toxicity Responses	
		Mean	Standard Error
	“Small variance”	7.5	0.063
	“Big variance”	6.8	0.15
	“Fixed”	7.5	0.062

Table 5.24 Number of Toxicity Responses when “ $\gamma=0.5$ ”

		Number of Toxicity Responses	
		Mean	Standard Error
	“Small variance”	4.3	0.062
	“Big variance”	6.2	0.096
	“Fixed”	4.1	0.06

Table 5.25 - Table 5.27 list the average proportions of toxicity responses and their standard errors across different designs and different underlying outcome models. The standard errors are about the same. Regardless of the underlying outcome models, no simulated clinical trials have early stopping (to be shown later), and thus the total number of enrolled patients in each simulated clinical trial is the same, equal to the maximum sample size. The observed trend

with respect to the design performance for the average proportion of toxicity responses is the same as those for the average number of toxicity responses.

Table 5.25 Proportion of Toxicity Responses when “ $\gamma = 1$ ”

		Proportion of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>			
	“Small variance”	0.206	0.002
	“Big variance”	0.278	0.004
	“Fixed”	0.201	0.002

Table 5.26 Proportion of Toxicity Responses when “ $\gamma = 2$ ”

		Proportion of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>			
	“Small variance”	0.251	0.002
	“Big variance”	0.228	0.005
	“Fixed”	0.25	0.002

Table 5.27 Proportion of Toxicity Responses when “ $\gamma = 0.5$ ”

		Proportion of Toxicity Responses	
		Mean	Standard Error
<b>Design</b>			
	“Small variance”	0.145	0.002
	“Big variance”	0.207	0.003
	“Fixed”	0.137	0.002

- Average percentage of patients treated at the true target dose level

Table 5.28 - Table 5.30 list the average percentages of patients treated below, at and above the true target dose level across different designs and different underlying outcome models. The true target dose level is 3 when “ $\gamma=1$ ”, 2 when “ $\gamma=2$ ”, and 5 when “ $\gamma=0.5$ ”. Adding a small random disturbance to the prior is found not to change the average percentage of patients very much except when  $\gamma=0.5$ . The average proportion of toxicity responses is the smallest in the “Fixed” design when “ $\gamma=1$ ” or “ $\gamma=0.5$ ”, and in the “Big variance” design when “ $\gamma=2$ ”.

Table 5.28 Percentage of Patients Treated when “ $\gamma=1$ ”

		<b>Below Target</b>	<b>At Target</b>	<b>Above Target</b>
<b>Design</b>	“Small variance”	45.1	47.3	7.6
	“Big variance”	30.8	38.7	30.5
	“Fixed”	46.2	47.4	6.4

Table 5.29 Percentage of Patients Treated when “ $\gamma=2$ ”

		<b>Below Target</b>	<b>At Target</b>	<b>Above Target</b>
<b>Design</b>	“Small variance”	13.7	79.4	6.9
	“Big variance”	38.4	51.2	10.3
	“Fixed”	13.7	79.3	7.1

Table 5.30 Percentage of Patients Treated when “ $\gamma=0.5$ ”

		<b>Below Target</b>	<b>At Target</b>	<b>Above Target</b>
<b>Design</b>	“Small variance”	84.1	15.9	0
	“Big variance”	29.2	70.8	0
	“Fixed”	89.1	10.9	0

- Proportion of simulations where the true target dose level is estimated correctly

Table 5.31 Proportion of Correct Estimation of the True Target when “ $\gamma = 1$ ”

		<b>Correct Estimation of the True Target</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Small variance”	0.64	0.596	0.682
	“Big variance”	0.572	0.527	0.616
	“Fixed”	0.616	0.572	0.659

Table 5.32 Proportion of Correct Estimation of the True Target when “ $\gamma = 2$ ”

		<b>Correct Estimation of the True Target</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Small variance”	0.926	0.899	0.947
	“Big variance”	0.6	0.556	0.643
	“Fixed”	0.916	0.888	0.939

Table 5.33 Proportion of Correct Estimation of the True Target when “ $\gamma = 0.5$ ”

		<b>Correct Estimation of the True Target</b>		
<b>Design</b>		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Small variance”	0.308	0.268	0.351
	“Big variance”	0.844	0.809	0.875
	“Fixed”	0.166	0.134	0.202

Table 5.31 - Table 5.33 list the proportions of correct estimation of the true target and their exact 95% confidence intervals across different designs and different underlying outcome

models. When “ $\gamma=1$ ” and “ $\gamma=2$ ”, the “Big variance” design has smallest proportion, while the other two designs have similar proportions; when “ $\gamma=0.5$ ”, the larger the variance of prior for  $\gamma$ , the larger proportion the corresponding design has.

- Proportion of early stopping

Table 5.34 - Table 5.36 list the proportions of early stopping and their exact 95% confidence intervals across different designs and different underlying outcome models. Under the settings in this experiment, since the true target dose level is within the testing dose levels, early stopping is not favorable. The average proportions of early stopping are all zero, which is probably because we use an informative prior with mode of 0.05 for  $p_s$ .

Table 5.34 Proportion of Early Stopping when “ $\gamma=1$ ”

		Early Stopping		
Design		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Small variance”	0	0	0.006
	“Big variance”	0	0	0.006
	“Fixed”	0	0	0.006

Table 5.35 Proportion of Early Stopping when “ $\gamma=2$ ”

		Early Stopping		
Design		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Small variance”	0	0	0.006
	“Big variance”	0	0	0.006
	“Fixed”	0	0	0.006

Table 5.36 Proportion of Early Stopping when “ $\gamma = 0.5$ ”

		Early Stopping		
Design		Proportion	95% CI	
			Lower Bound	Upper Bound
	“Small variance”	0	0	0.006
	“Big variance”	0	0	0.006
	“Fixed”	0	0	0.006

Results in Section 5.1.4.2 show that no designs perform uniformly best across different criteria and different underlying outcome models. Adding a small randomness to the fixed prior seems not to improve the design performance. All simulated clinical trials using different designs do not have early stoppings. The “Big variance” design has the smallest number and average proportion of toxicity responses when  $\gamma = 2$ , and significantly larger average percentage of patients treated at the true target dose level and proportion of the correct estimation of the true target dose level than the other two designs when  $\gamma = 0.5$ .

## 5.2 PATIENT HETEROGENEITY

Patients enrolled in Phase I cancer trials vary greatly in their current disease status, numbers and types of previous treatment, age, sex, genetic profile and many other factors that may impact their tolerance to the testing treatment. An example is capecitabine, a widely used fluoropyrimidine that requires dose reduction for elderly patients and those with mild to moderate renal insufficiency [83]. Ignoring patient heterogeneity in the Phase I trials may do harm to patients or recommend an either suboptimal or too toxic dose for future studies. In practice, however, probably due to the time and cost constraints, few Phase I clinical trials



account for the patient heterogeneity. We think a quantitative demonstration for the potential adverse effect that ignoring patient heterogeneity in a particular Phase I trial would help investigators to think about risks and benefits more clearly and make a good decision.

In this section, we will present two hypothetical examples showing how to assess via simulation the difference in the expected total personal utility and societal utility between trials accounting for patient heterogeneity and trials ignoring patient heterogeneity. One example considers a CRM design, and the other considers the standard “3+3” design. In both examples, we assume that there are two types of patient populations with respect to their toxicity tolerance to the testing drug; the population with better tolerance on average is called “Low Risk”, and the other population is called “High Risk”. We also assume that “Low Risk” and “High Risk” patients are mixed in 1:1 in a Phase I trial unless they are separated through measurements and treated differently in studies. Suppose patients have underlying dose thresholds for both efficacy and toxicity responses,  $\theta_E$  and  $\theta_T$ . The natural logarithm of dose thresholds follows a bivariate normal distribution. Expressions (17) - (18) give the dose thresholds’ distributions for “Low Risk”, and “High Risk” patient population respectively. The “L” subscript refers to “Low Risk”, and “H” refers to “High Risk”. Figure 5.19 displays the contour plot for their dose thresholds distributions.

$$\begin{pmatrix} \log(\theta_{EL}) \\ \log(\theta_{TL}) \end{pmatrix} \sim N \left( \begin{pmatrix} 2.71 \\ 3.72 \end{pmatrix}, \begin{bmatrix} 2.25 & 1.8 \\ 1.8 & 2.25 \end{bmatrix} \right) \quad (17)$$

$$\begin{pmatrix} \log(\theta_{EH}) \\ \log(\theta_{TH}) \end{pmatrix} \sim N \left( \begin{pmatrix} 1.1 \\ 1.64 \end{pmatrix}, \begin{bmatrix} 0.64 & 0.32 \\ 0.32 & 0.64 \end{bmatrix} \right) \quad (18)$$

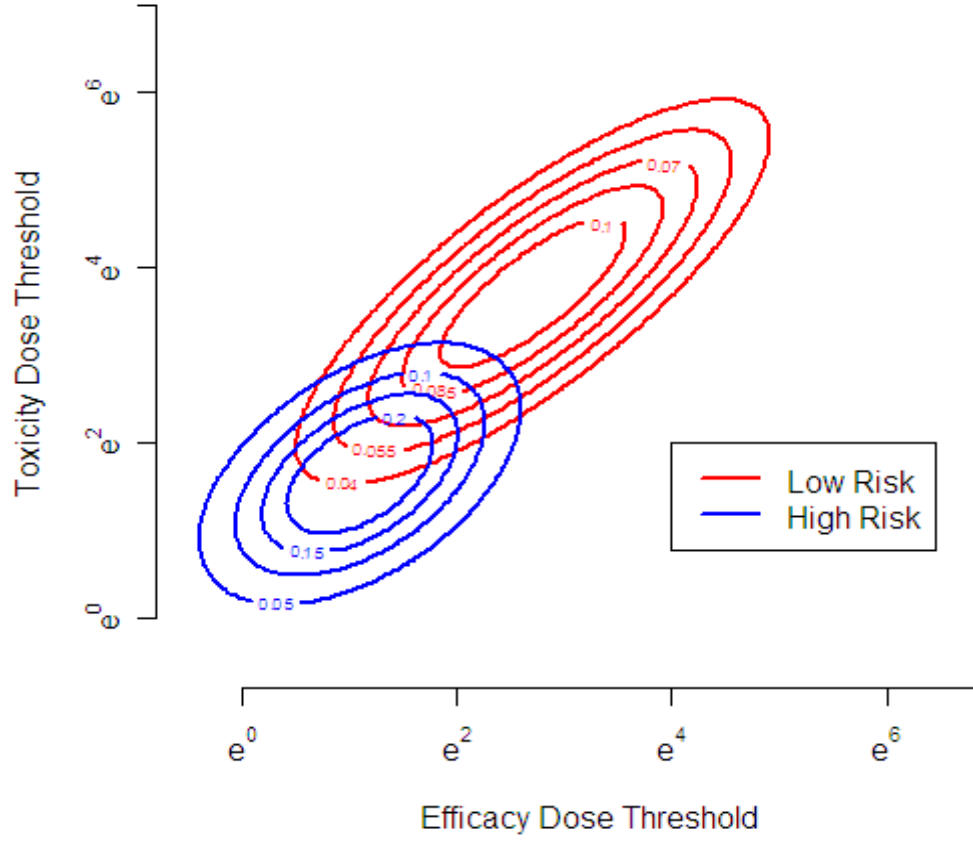


Figure 5.19 Contour Plot for Dose Thresholds Distribution

The dose thresholds distribution in the mixture patient population can be expressed by Equation (19), and the subscript “M” standards for “Mixture”.

$$\begin{pmatrix} \log(\theta_{TM}) \\ \log(\theta_{EM}) \end{pmatrix} \sim 0.5 \times N \left( \begin{pmatrix} 3.72 \\ 2.71 \end{pmatrix}, \begin{bmatrix} 2.25 & 1.8 \\ 1.8 & 2.25 \end{bmatrix} \right) + 0.5 \times N \left( \begin{pmatrix} 1.64 \\ 1.1 \end{pmatrix}, \begin{bmatrix} 0.64 & 0.32 \\ 0.32 & 0.64 \end{bmatrix} \right) \quad (19)$$

The outcome model used in the two examples can be described in Equation (20):

$$y_{hk} = \begin{cases} 1 & \theta_{hk} < z \\ 0 & \theta_{hk} > z \end{cases}, \quad h \in \{E, T\}; k \in \{L, H\}; z \in \mathcal{Z} \quad (20)$$

, where  $y_{hk}$  is the binary response for either efficacy or toxicity. The dose-toxicity and dose-efficacy curves for different patient populations are drawn in

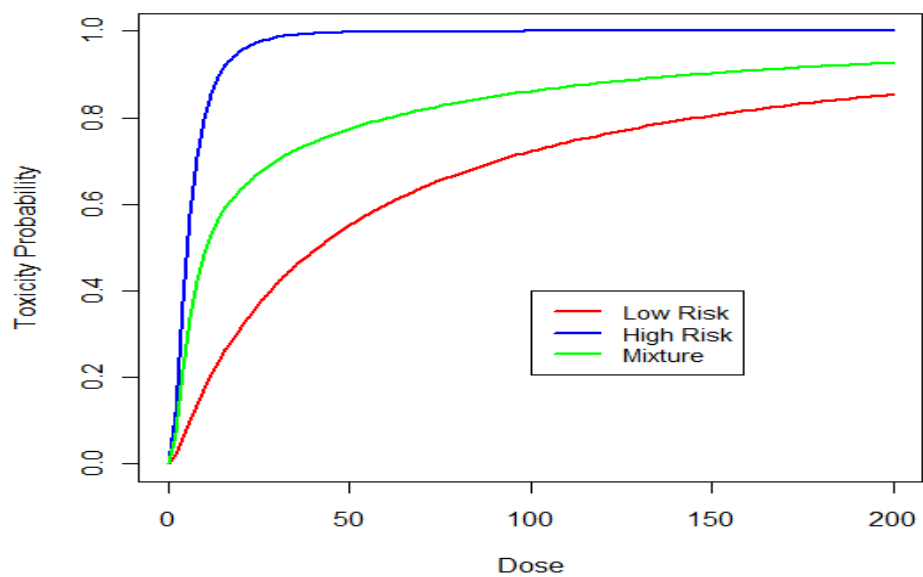


Figure 5.20 Dose-Toxicity Curve

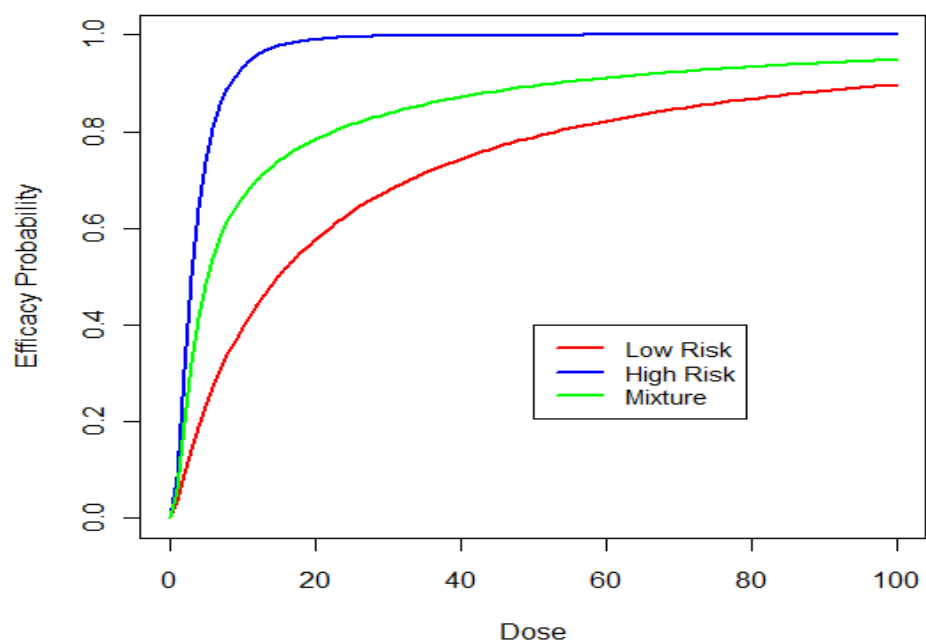


Figure 5.21 Dose-Efficacy Curve

We use the additive utility function to measure the utilities associated with four possible outcomes: TE, Te, tE, and te (even though the CRM and standard “3+3” designs only measure univariate toxicity outcomes, we can simulate the efficacy responses from the assumed true population and outcome models). To measure the expected societal utility, we consider the utility function described in Section 3.2.2. Suppose the current standard treatment has  $(P_E, P_T) = (0.15, 0.1)$ , and investigators think if the testing drug at the chosen RP2D has  $(P_E, P_T) = (0.05, 0)$ , the societal utility gained from this trial is 0. We also assume the societal utility is 0 if no RP2D is chosen at the end of a trial.

Table 5.37 Schemes in Consideration

Scheme #	Scheme Name	Tier Doses
1	“Mixed, assumed low risk”	3, 6.6, 9, 15, 21.1
2	“Mixed, assumed high risk”	0.91, 1.82, 3, 4.56, 6.43
3	“Separate”	For low risk: 3, 6.6, 9, 15, 21.1 For high risk: 0.91, 1.82, 3, 4.56, 6.43

We compare three schemes, the first and second schemes do not distinguish two patient groups and have Phase I trials with mixture patient population. In the first scheme, the investigators assume the population is low risk and they test tier doses:  $\mathcal{X}_L = \{3, 6.6, 9, 15, 21.1\}$ ; the investigators in the second scheme assume the population is high risk and they test tier doses:  $\mathcal{X}_H = \{0.91, 1.82, 3, 4.56, 6.43\}$ ; the third scheme distinguishes two patient groups, and run two separate Phase I trials, the tier doses  $\mathcal{X}_L$  is used for the trial with pure low risk population, and the tier doses  $\mathcal{X}_H$  is used for the trial with pure high risk

population. We run 1000 simulations for each scheme. We will look at comparison results with respect to expected total personal utility and expected societal utility under the CRM and the standard “3+3” design in the following two sections.

### 5.2.1 The CRM

The three schemes using the CRM share the following common design parameters:

- Number of testing dose levels: 5
- The starting dose level is the first dose level
- Single-stage
- One patient per cohort
- Target toxicity rate: 0.25
- Early Stopping: the posterior toxicity probability at the first dose level is larger than  $(0.1 + \text{target}) = 0.35$
- $\gamma$ : fixed at 1
- Prior for  $p_s$ : Beta(1.47,10)
- Initial toxicity probability guesses: 0.05, 0.12, 0.25, 0.46, 0.68

Besides the difference in the testing tier doses, the three schemes use different maximum sample size, as illustrated in Table 5.38.

The evaluation results in Table 5.39 and Table 5.40 show that assuming patients are low risk in a trial when the actual patient population is mixed has around eight standard errors larger expected total personal utility, and slightly larger expected societal utility than assuming high risk patient population. No early stopping is observed in all the simulated trials for the three

schemes, thus on average, two separate trials in the third scheme each only enroll half the patients of the trial in the first two scheme. Since these two separate trials are run independently from each other and they together enroll the same number of patients as the other two trials, it would be reasonable to add their expected utilities together and compare with the expected utilities in the other two trials. Obviously, trials together in the third scheme have the largest expected total personal utility and societal utility.

Table 5.38 Different Maximum Sample Sizes in the CRM

<b>Scheme Name</b>	<b>Maximum Sample Size</b>
“Mixed, assumed low risk”	32
“Mixed, assumed high risk”	32
“Separate”	Each trial: 16

Table 5.39 Expected Total Personal Utility Using the CRM

<b>Scheme Name</b>	$EU_p$	<b>Standard Error</b>
“Mixed, assumed low risk”	6.2	0.088
“Mixed, assumed high risk”	5.5	0.087
“Separate”	For low risk: 3.3 For high risk: 3.3	For low risk: 0.058 For high risk: 0.069

Table 5.40 Expected Societal Utility Using the CRM

<b>Scheme Name</b>	$EU_s$	<b>Standard Error</b>
“Mixed, assumed low risk”	0.174	0.001
“Mixed, assumed high risk”	0.172	0.001
“Separate”	For low risk: 0.22 For high risk: 0.229	For low risk: 0.001 For high risk: 0.002

### 5.2.2 The standard “3+3” design

The following two tables summarize the evaluation results for the three schemes when they are applied to the standard “3+3” design. Unlike the CRM designs in Section 0, we do not add the maximum sample size restriction to the standard “3+3” design. Results show that clinical trials that run separately for different patient groups have much larger expected total personal utility and societal utility than clinical trials that do not run separately.

Table 5.41 Expected Total Personal Utility Using the “3+3” Design

<b>Scheme Name</b>	$EU_p$	<b>Standard Error</b>
“Mixed, assumed low risk”	2.2	0.063
“Mixed, assumed high risk”	2.6	0.064
“Separate”	For low risk: 3.4 For high risk: 3.1	For low risk: 0.07 For high risk: 0.078

Table 5.42 Expected Societal Utility Using the “3+3” Design

<b>Scheme Name</b>	$EU_s$	<b>Standard Error</b>
“Mixed, assumed low risk”	0.135	0.002
“Mixed, assumed high risk”	0.114	0.002
“Separate”	For low risk: 0.153 For high risk: 0.187	For low risk: 0.003 For high risk: 0.003

### 5.2.3 Conclusions

We present a way to quantitatively measure the expected total personal utility and societal utility from a particular clinical trial through simulations. The two hypothetical examples both demonstrate that ignoring patient heterogeneity in the CRM and the standard “3+3” design can significantly decrease the expected total personal utility and societal utility.



## 6.0 CONCLUSIONS AND FUTURE WORK

CT simulation has become an important and useful tool to improve the efficiency and accuracy of drug development. This dissertation attempts to improve, in particular, the design evaluation process in three directions.

First, we develop an open-source highly flexible and extendible simulation experiment platform for evaluating CT designs. The S4 system of classes and methods is utilized. Using object-oriented programming provides extensibility through careful, clear interface specification; using R, an open-source widely-used statistical language, makes the application extendible by the people who design CTs: biostatisticians. Action queue-based simulation framework mimics a real CT setting, and adaptively adds actions to the queue, as needed, guided by current state of information. Currently, name matching mechanism is used to check interoperability among the objects.

Second, we propose flexible criteria for evaluating Phase I trial designs by assessing through CT simulation the expected total personal utility, societal utility and total utility. The expected utility paradigm of Bayesian decision theory is well suited to helping CT designers with incorporating prior information, and providing a comprehensive design evaluation.

Third, we present several examples using the platform to investigate important questions in clinical trial designs. Specifically, we look at the logit model in the continual reassessment method (CRM), choices of parameterization and prior distribution for its model parameters. We

demonstrate that using easily interpretable parameters would help to choose the number of free parameters in a model, and to set up sensible prior distributions that genuinely reflect the investigators' prior beliefs. We also look at the effect of patient heterogeneity on the performance of the standard "3+3" design and the CRM. We observe that ignoring patient heterogeneity in the CRM and the standard "3+3" design can significantly decrease the expected total personal utility and societal utility.

To facilitate the use of platform, we have developed an R package called CTDesignExplorer based on the source code as of April 30<sup>th</sup>, 2010. It is intended to be both open-source and open-development. Future work will include setting up requirements for contributors, publishing CTDesignExplorer, project registration on R-Forge to facilitate collaborative software development, adapting code for parallel processing and building a user-friendly graphical user interface (GUI).

## BIBLIOGRAPHY

1. Hale, M., et al., *Clinical Trial Simulation as a Tool for Increased Drug Development Efficiency*. Applied Clinical Trials, 1996. 5: p. 35-40.
2. Holford, N.H.G., et al., *Simulation of Clinical Trials*. Annual Review of Pharmacology and Toxicology, 2000. 40: p. 209-234.
3. Lockwood, P., et al., *Application of clinical trial simulation to compare proof-of-concept study designs for drugs with a slow onset of effect; an example in Alzheimer's disease*. Pharmaceutical Research, 2006. 23(9): p. 2050-2059.
4. Santen, G., et al., *From Trial and Error to Trial Simulation. Part 1: The Importance of Model-Based Drug Development for Antidepressant Drugs*. Clinical Pharmacology & Therapeutics, 2009. 86(3): p. 248-254.
5. The Open Source Initiative, *The Open Source Definition*, 1998. November 1<sup>st</sup>, 2010 <<http://www.opensource.org/osd.html>>.
6. O'Quigley, J., M. Pepe, and L. Fisher, *Continual Reassessment Method: A Practical Design for Phase I Clinical Trials in Cancer*. Biometrics, 1990. 46: p. 33-48.
7. Braun, T.M., *The Bivariate Continual Reassessment Method: Extending the CRM to Phase I Trials of Two Competing Outcomes*. Controlled Clinical Trials, 2002. 23: p. 240-256.
8. Goodman, S.N., M.L. Zahurak, and S. Piantadosi, *Some Practical Improvements in the Continual Reassessment Method for Phase I Studies*. Statistics in Medicine, 1995. 14: p. 1149-1161.
9. Moller, S., *An Extension of the Continual Reassessment Methods Using a Preliminary Up-and-Down Design in a Dose Finding Study in Cancer Patients, in Order to Investigate a Greater Range of Doses*. Statistics in Medicine, 1995. 14: p. 911-922.
10. Piantadosi, S., J.D. Fisher, and S. Grossman, *Practical Implementation of a Modified Continual Reassessment Method for Dose-finding Trials*. Cancer Chemotherapeutics and Pharmacology, 1998. 41: p. 429-436.

11. Faries, D., *Practical Modifications of the Continual Reassessment Method for Phase I Cancer Clinical Trials*. Journal of Biopharmaceutical Statistics, 1994. 4: p. 147-164.
12. Heyd, J.M. and B.P. Carlin, *Adaptive Design Improvements in the Continual Reassessment Method for Phase I Studies*. Statistics in Medicine, 1999. 18: p. 1307-1321.
13. Ishizuka, N. and Y. Ohashi, *The Continual Reassessment Method and its Applications: a Bayesian Methodology for Phase I Cancer Clinical Trials*. Statistics in Medicine, 2001. 20: p. 2661-2681.
14. Cheung, Y.-K. and R. Chappell, *Sequential Designs for Phase I Clinical Trials with Late-Onset Toxicities*. Biometrics, 2000. 56: p. 1177-1182.
15. O'Quigley, J. and L.Z. Shen, *Continual Reassessment Method: A Likelihood Approach*. Biometrics, 1996. 52: p. 673-684.
16. Gerke, O. and H. Siedentop, *Optimal Phase I Dose-escalation Trial Designs in Oncology—A Simulation Study*. Statistics in Medicine, 2008. 27: p. 5329-5344.
17. Gerke, O. and S. H., *Authors' Rejoinder to 'Dose-escalation Designs in Oncology: ADEPT and the CRM'*. Statistics in Medicine, 2008. 27: p. 5354-5355.
18. Shu, J. and J. O'Quigley, *Dose-escalation Designs in Oncology: ADEPT and the CRM*. Statistics in Medicine, 2008. 27: p. 5345-5353.
19. Shen, L.Z. and J. O'Quigley, *Using A One-parameter Model to Sequentially Estimate the Root of A Regression Function*. Computational Statistics & Data Analysis, 2000. 34: p. 357-369.
20. Neuenschwander, B., M. Branson, and T. Gsponer, *Critical Aspects of the Bayesian Approach to Phase I Cancer Trials*. Statistics in Medicine, 2008. 27: p. 2420-2439.
21. Eisenhauer, E.A., et al., *Phase I Clinical Trial Design in Cancer Drug Development*. Journal of Clinical Oncology, 2000. 18: p. 684-692.
22. Dent, S.F. and E.A. Eisenhauer, *Phase I Trial Design: Are New Methodologies Being Put into Practice?* Annals of Oncology, 1996. 7: p. 561-566.
23. Mahmood, I., *Interspecies Scaling of Maximum Tolerated Dose of Anticancer Drugs: Relevance to Starting Dose for Phase I Trials*. American Journal of Therapeutics, 2001. 8: p. 109-116.
24. Koyfman, S.A., et al., *Risks and Benefits Associated with Novel Phase I Oncology Trial Designs*. Cancer, 2007. 110: p. 1115-1124.
25. Levine, R.J., *Ethics and Regulation of Clinical Research*. 2nd ed. 1986: Yale University Press.

26. Fox, R.M., *Situations Creating Ethical Stress*. The Medical Journal of Australia, 1981. 1: p. 162-163.
27. Lipsett, M.B., *On the Nature and Ethics of Phase I Clinical Trials of Cancer Chemotherapies*. Journal of the American Medical Association, 1982. 248: p. 941-942.
28. Markman, M., *The Ethical Dilemma of Phase I Clinical Trials*. CA: A Cancer Journal for Clinicians, 1986. 36: p. 367-369.
29. Sass, H.M., *Ethical Considerations in Phase I Clinical Trials*. Onkologie, 1990. 13: p. 85-88.
30. Kodish, E., et al., *Ethical Issues in Phase I Oncology Research: A Comparison of Investigators and Institutional Review Board Chairpersons*. Journal of Clinical Oncology, 1992. 10: p. 1810-1816.
31. Agrawal, M. and E.J. Emanuel, *Ethics of Phase I Oncology Studies: Reexamining the Arguments and Data*. Journal of the American Medical Association, 2003. 290: p. 1075-1082.
32. Wang, M., *An Adaptive Bayesian Approach to Jointly Modeling Response and Toxicity in Phase I Dose-Finding Trials*, in Biostatistics. 2007, University of Pittsburgh: Pittsburgh.
33. Gasparini, M. and J. Eisele, *A Curve-Free Method for Phase I Clinical Trials*. Biometrics, 2000. 56: p. 609-615.
34. Mukhopadhyay, S., *Bayesian Nonparametric Inference on the Dose Level with Specified Response Rate*. Biometrics, 2000. 56: p. 220-226.
35. Whitehead, J. and H. Brunier, *Bayesian Decision Procedures for Dose Determining Experiments*. Statistics in Medicine, 1995. 14: p. 885-893.
36. Thall, P.F., H.Q. Nguyen, and E.H. Estey, *Patient-Specific Dose Finding Based on Bivariate Outcomes and Covariates*. Biometrics, 2008. 64: p. 1126-1136.
37. Edler, L. and I. Burkholder, *Overview of Phase I Trials*, in Handbook of Statistics in Clinical Oncology, J. Crowley and D.P. Ankerst, Editors. 2006, Taylor & Francis Group, LLC: Boca Raton. p. 3-29.
38. Garrett-Mayer, E., *The Continual Reassessment Method for Dose-Finding Studies: A Tutorial*. Clinical Trials, 2006. 3: p. 57-71.
39. Dragalin, V. and V.V. Fedorov, *Adaptive Designs for Dose-Finding Based on Efficacy-Toxicity Response*. Journal of Statistical Planning and Inference, 2006. 136: p. 1800-1823.
40. Potter, D.M., *Phase I Studies of Chemotherapeutic Agents in Cancer Patients: A Review of the Designs*. Journal of Biopharmaceutical Statistics, 2006. 16: p. 579-604.

41. Chevret, S., ed. *Statistical Methods for Dose-Finding Experiments*. 2006, John Wiley & Sons Ltd: Chichester.
42. Ratain, M.J., et al., *Statistical and Ethical Issues in the Design and Conduct of Phase I and II Clinical Trials of New Anticancer Agents*. Journal of the National Cancer Institute, 1993. 85: p. 1637-1643.
43. Smith, T.L., et al., *Design and Results of Phase I Cancer Clinical Trials: Three-Year Experience at M.D. Anderson Cancer Center*. Journal of Clinical Oncology, 1996. 14: p. 287-295.
44. Zohar, S. and S. Chevret, *Recent Developments in Adaptive Designs for Phase I/II Dose-Finding Studies*. Journal of Biopharmaceutical Statistics, 2007. 17: p. 1071-1083.
45. Rosenberger, W.F. and L.M. Haines, *Competing Designs for Phase I Clinical Trials: A Review*. Statistics in Medicine, 2002. 21: p. 2757-2770.
46. Ting, N., *Dose Finding in Drug Development*. 2006, New York: Springer.
47. Storer, B.E., *Design and Analysis of Phase I Clinical Trials*. Biometrics, 1989. 45: p. 925-937.
48. Geller, N.L., *Design of Phase I and II Clinical Trials in Cancer: A Statistician's View*. Cancer Investigation, 1984. 2: p. 483-491.
49. Yuan, Z., R. Chappell, and H. Bailey, *The Continual Reassessment Method for Multiple Toxicity Grades: A Bayesian Quasi-Likelihood Approach*. Biometrics, 2007. 63: p. 173-179.
50. Liu, G., W.F. Rosenberger, and L.M. Haines, *Sequential Designs for Ordinal Phase I Clinical Trials*. Biometrical Journal. 51: p. 335-347.
51. Ivanova, A. and S.H. Kim, *Dose Finding for Continuous and Ordinal Outcomes with a Monotone Objective Function: A Unified Approach*. Biometrics, 2009. 65: p. 307-315.
52. Potthoff, R.F. and S.L. George, *Flexible Phase I Clinical Trials: Allowing for Nonbinary Toxicity Response and Removal of Other Common Limitations*. Statistics in Biopharmaceutical Research, 2009. 1: p. 213-228.
53. Yin, G. and Y. Yuan, *Bayesian Model Averaging Continual Reassessment Method in Phase I Clinical Trials*. Journal of the American Statistical Association, 2009. 104: p. 954-968.
54. Babb, J.S., A. Rogatko, and S. Zacks, *Cancer Phase I Clinical Trials: Efficient Dose Escalation with Overdose Control*. Statistics in Medicine, 1998. 17: p. 1103-1120.
55. Chu, P.-L., Y. Lin, and W.J. Shih, *Unifying CRM and EWOC Designs for Phase I Cancer Clinical Trials*. Journal of Statistical Planning and Inference 2008. 139: p. 1146-1163.

56. Huang, B. and R. Chappell, *Three-Dose-Cohort Designs in Cancer Phase I Trials*. Statistics in Medicine, 2008. 27: p. 2070-2093.
57. Bekele, B.N., et al., *Monitoring Late-Onset Toxicities in Phase I Trials using Predicted Risks*. Biostatistics, 2008. 9: p. 442-457.
58. Fan, S.K. and Y.-G. Wang, *Designs for Phase I Clinical Trials with Multiple Courses of Subjects at Different Doses*. Biometrics, 2007. 63: p. 856-864.
59. Leung, D. and Y.-G. Wang, *Isotonic Designs for Phase I Trials*. Controlled Clinical Trials, 2001. 22: p. 126-138.
60. Cheung, Y.K., *Sequential Implementation of Stepwise Procedures for Identifying the Maximum Tolerated Dose*. Journal of the American Statistical Association, 2007. 102: p. 1448-1461.
61. Zandvliet, A.S., et al., *Two-stage Model-based Clinical Trial Design to Optimize Phase I Development of Novel Anticancer Agents*. Investigational New Drugs, 2010. 28: p. 61-75.
62. Skolnik, J.M., et al., *Shortening the Timeline of Pediatric Phase I Trials: The Rolling Six Design*. Journal of Clinical Oncology, 2008. 26: p. 190-195.
63. Fan, S.K., A.P. Venook, and Y. Lu, *Design Issues in Dose-Finding Phase I Trials for combinations of Two Agents*. Journal of Biopharmaceutical Statistics, 2009. 19: p. 509-523.
64. Yin, G. and Y. Yuan, *A Latent Contingency Table Approach to Dose Finding for Combinations of Two Agents*. Biometrics, 2009. 65: p. 866-875.
65. Yin, G. and Y. Yuan, *Bayesian Dose Finding in Oncology for Drug Combinations by Copula Regression*. Journal of the Royal Statistical Society, Series C, 2009. 58: p. 211-224.
66. Murtaugh, P.A. and L.D. Fisher, *Bivariate Binary Models of Efficacy and Toxicity in Dose-Ranging Trials*. Communications in Statistics, Part A, 1990. 19: p. 2003-2020.
67. Thall, P.F. and J.D. Cook, *Dose-Finding Based on Efficacy-Toxicity Trade-Offs*. Biometrics, 2004. 60: p. 684-693.
68. Ivanova, A., et al., *An Adaptive Design for Identifying the Dose with the Best Efficacy/Tolerability Profile with Application to a Crossover Dose-Finding Study*. Statistics in Medicine, 2009. 28: p. 2941-2951.
69. Dragalin, V., V.V. Fedorov, and Y. Wu, *Two-Stage Design for Dose-Finding that Accounts for Both Efficacy and Safety*. Statistics in Medicine, 2008. 27: p. 5156-5176.

70. Mandrekar, S.J., Y. Cui, and D.J. Sargent, *An Adaptive Phase I Design for Identifying a Biologically Optimal Dose for Dual Agent Drug Combinations*. *Statistics in Medicine*, 2007. 26: p. 2317-2330.
71. Zhang, W., D.J. Sargent, and S. Mandrekar, *An Adaptive Dose-Finding Design Incorporating Both Toxicity and Efficacy*. *Statistics in Medicine*, 2006. 25: p. 2365-2383.
72. Huang, X., et al., *A Parallel Phase I/II Clinical Trial Design for Combination Therapies*. *Biometrics*, 2007. 63: p. 429-436.
73. Wang, M. and R. Day, *Adaptive Bayesian Design for Phase I Dose-finding Trials Using A Joint Model of Response and Toxicity*. *Journal of Biopharmaceutical Statistics*, 2010. 20: p. 125-144.
74. Ahn, C., *An Evaluation of Phase I Cancer Clinical Trial Designs*. *Statistics in Medicine*, 1998. 17: p. 1537-1549.
75. Lin, Y. and W.J. Shih, *Statistical Properties of the Traditional Algorithm-Based Designs for Phase I Cancer Clinical Trials*. *Biostatistics*, 2001. 2: p. 203-215.
76. Storer, B.E., *An Evaluation of Phase I Clinical Trial Designs in the Continuous Dose-Response Setting*. *Statistics in Medicine*, 2001. 20: p. 2399-2408.
77. Paoletti, X., J. O'Quigley, and J. Maccario, *Design Efficiency in Dose Finding Studies*. *Computational Statistics & Data Analysis*, 2004. 45: p. 197-214.
78. Chambers, J.M., *Software for Data Analysis*. 2008, New York: Springer Science+Business Media, LLC.
79. Jones, C.L. and E. Holmgren, *An Adaptive Simon Two-Stage Design for Phase 2 Studies of Targeted Therapies*. *Contemporary Clinical Trials*, 2007. 28: p. 654-661.
80. Raymond, E.S., *The Cathedral and the Bazaar*. *Knowledge, Technology & Policy*, 1999. 12: p. 23-49.
81. R-Forge Administration and Development Team, *R-Forge User's Manual, Beta*. 2009. November 1<sup>st</sup>, 2010 <<http://r-forge.r-project.org/>>.
82. Eddelbuettel, D. *High-Performance and Parallel Computing with R*, 2010. November 1<sup>st</sup>, 2010 <<http://cran.r-project.org/web/views/HighPerformanceComputing.html>>.
83. Poole, C., et al., *Effect of Renal Impairment on the Pharmacokinetics and Tolerability of Capecitabine (Xeloda) in Cancer Patients*. *Cancer Chemotherapy and Pharmacology*, 2002. 49: p. 225-234.